RDTR NO. 282

EMPLOYEE PERFORMANCE EVALUATION AND REVIEW: A SUMMARY OF THE LITERATURE

PREPARED BY

APPLIED SCIENCES DEPARTMENT

NAVAL AMMUNITION DEPOT, CRANE, INDIANA

NAVAL AMMUNITION DEPOT
CRANE, INDIANA

RDTR No. 282
9 August 1974

EMPLOYEE PERFORMANCE EVALUATION AND REVIEW: A SUMMARY OF THE LITERATURE

BY

MARK S. SANDERS

JAMES M. PEAY

RELEASED BY:

James M. Peay

JAMES M. PEAY, Manager
Human Engineering Division
Applied Sciences Department

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RDTR No. 282	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Employee Performance Evaluation and Review: A Summary of the Literature		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR*(s)* MARK S. SANDERS		8. CONTRACT OR GRANT NUMBER*(s)*
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Ammunition Depot Crane, Indiana 47522		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE 9 August 1974
		13. NUMBER OF PAGES 89
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)*
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Statement A. Distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This report is a survey of literature dealing with employee evaluation and review techniques. The literature reviewed comes primarily from psychological and professional business journals. The report is organized around the decisions which must be made in order to implement a performance appraisal program.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

TABLE OF CONTENTS

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report is a survey of the literature dealing with employee evaluation and review techniques. The literature reviewed comes primarily from psychological and professional business journals. The report is organized around the decisions which must be made in order to implement a performance appraisal program.

There is no question about whether evaluations should be made. The question is: how are we going to evaluate? We can decide to indulge in making capricious judgments about people with each evaluator given full rein to his own standards and biases or we can choose to evaluate people according to an organized and systematic procedure which attempts to set up common standards of judgment which all evaluators can apply uniformly and without bias. Obviously, the latter choice is the wiser one.

Performance appraisal is a complex and delicate matter that cannot be taken lightly. An effective performance appraisal system does not "just happen", it must be carefully planned and continuously monitored. There is no cheap effective system. You get what you are willing to pay for.

It makes only good sense to inventory what skills are available within the work force, develop potential where it exists, and use the human resource to its fullest potential. To do this, performance appraisal is a must.

The first and most critical step in developing a performance appraisal system is to formulate its aims and purposes. The specific

purposes for which performance appraisal have been used can be grouped under employee development or administrative actions. It is recommended to limit the purposes of a system to but a few. A major reason for the failure of appraisal systems is that they tried to do too much. Do not design a system to both foster employee development and supply information for administrative actions. Keep the two separate.

Employees should be told how they were evaluated. This is a must if employee development is the goal. Even if administrative action was the goal, employees still want to know how they stand with their supervisor. Supervisors must be trained to conduct performance reviews with their subordinates. The best approach to the performance review is a problem-solving approach patterned after management by objectives. With new or inexperienced employees, there is some evidence that a "tell and sell" approach may be more effective. Failure to build the appraisal program in the light of the demands of the post-appraisal interview may result in its emasculation.

The decision regarding who should actually do the appraisal is extremely important to an organization. The biases of the rater will in large part determine the future philosophy of the organization. Traditionally, the employee's immediate supervisor has been the chief appraiser of his performance, but recent writings suggest other alternatives such as peer ratings, self-ratings and for supervisors--subordinate ratings. The most important requirement for a rater is that he is familiar with the person he is evaluating and the job requirements of the incumbent's position. Peer ratings are perhaps the purest and best measure of leadership available. Self-appraisal

is usually incorporated in a management by objectives approach to appraisal. For supervisors, the use of subordinate ratings appears valuable.

A look at the history of industrial performance appraisal shows a distinct evolution from a dependence or personality trait rating through an almost equal passion for evaluating observable behavior to the most current compromise position of evaluating personality traits using observable behavior for definition.

Specific rating and ranking techniques are discussed in Chapter 7 following a discussion of the sources of distortion in rating procedures (e.g., leniency and halo errors). It is concluded that the best technique for employee development is the management by objectives approach. For administrative action, the field review method is considered the best.

Probably the major cause of failure in performance appraisal systems is the lack of training given the rater. It has been found that training increases the validity of the evaluation, its reliability, and reduces errors of leniency and halo. A good appraisal training program may require several training sessions and work shops. The three main areas which must be covered are: the value and importance of the program, how to make evaluations, and how to conduct the appraisal discussion with the ratee.

The last chapter, Chapter 9, recommends that a management by objectives performance review system be implemented for all employees, supervisors, and managers within the Applied Sciences Department of

NAD Crane. The recommendations include an evaluation of the current management by objectives program, appointment of a committee to implement the program, extensive training of supervisors, and an evaluation of the performance review program during the first year. In addition, it is recommended that subordinates evaluate their supervisors. Elaborate precautions would be taken to insure the anonymity of the subordinate, and no one but the supervisor would know how he was evaluated.

1. INTRODUCTION

Emperors of the Wei Dynasty (221-265 AD) were aided by "Imperial Raters" who appraised the performance of the members of the official family (Whisler and Harper, 1962). Today, in industry and government agencies, "Imperial Raters" still go about their task of appraising the performance of the official and not-so-official family. Despite its early beginnings, it was not until the 1800's that government in the United States started appraising performance. Industry didn't really get around to it until World War I.

Since then, thousands of articles and books have been written about performance appraisal. Four major trends have been noted over the years (Sloan and Johnson, 1968). First, the scope of performance appraisal has grown. In the early days (1920-1940) the emphasis was on appraisal of personality traits, the shift now is toward behavior and considering each man's contribution to the organization (i.e., management by objectives). Second, a trend has been to use appraisal more as a basis for employee development than for administrative actions such as salary administration. Third, formal evaluation of non-supervisory personnel has been decreasing and formal evaluation of supervisory personnel has been increasing. Fourth, there has been a growth in the psychometric sophistication of the method of appraisal.

Before continuing, it would be beneficial to discuss what is meant by the term performance appraisal. As used here, it is rubric which encompasses formal systems or program in which the work performance

of rank and file employees or supervisors is assessed. We are talking here about industrial and government jobs. Zeitlein (1969) makes an important distinction between performance reviews and performance evaluations. As used here, performance appraisal includes both review and evaluation. Performance evaluation is essentially quantitative and aims to supply management with information concerning the work force assets and liabilities. Here the emphasis is on rating scales. Performance review concentrates on the diagnosis of faults and correction through personal development. Here, the emphasis is on the performance review discussions between the employee and the reviewer (typically his immediate supervisor). The dichotomy is not always clear in practice. Often performance evaluations serve as the basis for a performance review. It is interesting to note that two major sources of literature, psychology and business, seem to divide in their emphasis. Psychological literature seems more concerned with performance evaluation, dwelling on rating form construction, bias, reliability, and validity. Business literature seems more concerned with performance review, dwelling primarily on how to conduct review discussions--the do's and don'ts.

This report is organized around the decisions which must be made in order to implement a performance appraisal (review and/or evaluation) program. Each chapter is a question which must be answered. Relevant literature is reviewed concerning each question. The questions roughly correspond to the order in which they should be answered. The answer to the first question will in part determine the answer to the next

question and so on. The basic questions were gleamed from an article published by Bittner in 1948. Even today, it serves as an excellent format around which to organize a review of performance appraisal.

The last chapter presents recommendations for implementing a performance review system within the Applied Sciences Department of the Naval Ammunition Depot, Crane, Indiana.

2. TO EVALUATE OR NOT TO EVALUATE . . . . .

That is not the question. It is human nature to compare and form judgments about people and things. It is virtually impossible for a manager, supervisor, foreman or employee to work with another person without forming a judgment about him. More importantly, that judgment will influence his actions toward that person and his subsequent perceptions of, and attitudes about, that person. The question then is not whether to evaluate or not, we have no choice, we will evaluate. The question is: "how are we going to evaluate"? We can decide to indulge in making capricious judgments about people, with each evaluator given full rein to his own standards and biases or we can choose to evaluate people according to an organized and systematic procedure which attempts to set up common standards of judgment which all evaluators can apply uniformly and without bias. Obviously, the latter choice is the wiser one.

Some people, however, have questioned or at least raised objections, to systematic performance appraisal systems. Firstly, it has been pointed out that supervisors are reluctant to use the system (Rowe, 1964; McGregor, 1957). McGregor (1957) has been one of the most outspoken critics of traditional performance appraisal because he feels it requires supervision to act as judges and they don't like to "play God" or treat people like products on an inspection line. Burke (1972) argues that the performance appraisal process takes time and has no pay-off for the manager. Many supervisors dislike the face-to-face confrontations required in performance reviews (Burke,

1972; Rowe, 1964; Thompson, 1969). Usually the supervisor is not trained in the skills required for effective performance review. In fact, the people orientation necessary for effective performance review is often the very thing that is considered least when promoting a man to supervisor.

Secondly, there have been reports of negative reactions among employees toward appraisal systems (VanZelst & Kerr, 1953; Meyer, Kay, and French, 1965). Apparently, the employees have a higher evaluation of themselves than do their supervisors (Barrett, 1966; Springer, 1953; Thornton, 1968; Parker, et. al., 1959, Rothaus, et. al., 1965) and the employee believes the supervisor will rate him higher than the supervisor actually does (Parker, et. al., 1959). This leads to a rather deflating experience for the employee and leads to negative feelings toward the appraisal system and often management as well.

Thirdly, some have questioned whether performance appraisal and review actually leads to improvements in employee performance (Burke, 1972; Culbreth, 1971; Meyer, Kay, and French, 1965). Some reasons for the lack of improvement are inadequate follow-up, too much time between evaluations and reviews, and lack of specific suggestions for improving performance.

In short, performance appraisal does not always work, it sometimes engenders negative reactions from both supervisors and employees and does not always lead to improvements in employee performance. It is necessary to note, however, that every author who has criticized traditional performance appraisal has ended with suggestions for im-

proving systematic performance appraisal systems, never has anyone suggested eliminating it. A close look at the criticisms reveal that the problems are due to the inadequate implementation of the systems rather than in any inherent fault with systematic appraisal itself. To suggest that, because systematic appraisal has not always worked, it should not be used is exemplary of the baby and the bath water over-reaction.

In fact, probably for every unsuccessful appraisal program one can find another similar one that has been successful. Studies have found positive attitudes toward performance appraisal (Clingenpeel, 1962) and improvements in employee performance following review (Meyer, Kay, and French, 1965).

A study by Spriegel (1962) surveyed 567 companies and found 257 having an appraisal program for executives and 343 having a program for foremen and below. One hundred eighty-four (184) companies had discontinued foremen and below appraisal, 256 had discontinued executive appraisal. The most frequent reason given for dropping a program was that the time required for appraisal became excessive.

Performance appraisal is a complex and delicate matter that cannot be taken lightly by anyone involved, from top management to the lowest level employee. An effective performance appraisal system does not "just happen", it must be carefully planned and continuously monitored. There is no cheap effective system. You get what you are willing to pay for.

It is interesting that business and government spend inordinate amounts of time and devote large numbers of men to inventory their capital resources (money, raw materials, machinery) yet are not willing to devote the same time and energy to inventory their human resources. Many executives fail to appreciate the importance of the human resources of their company. Humans are a major monetary expense--they furnish the know how and skills to run the company and can cause or compensate for inefficiencies. It makes only good sense to inventory what skills are available within the work force, develop potential where it exists, and use the human resource to its fullest potential. To do this, performance appraisal is a must.

Given, then, that systematic performance appraisal is the only intelligent choice, as appraisal will occur with or without a system, the next question is what do you want the system to do. That is, what are the goals of the system.

3. WHAT ARE THE AIMS AND PURPOSES?

The first, and most critical, step in developing a performance appraisal system is to formulate its aims and purposes. This must be carefully considered because the outcome will determine in great measure the form the system will take. Hayden (1973) lists five areas which are determined by the aims and purposes of the system: what is going to be appraised, the technique used to appraise, the supervisor's role in the process, the use and dissemination of the appraisal information, and the proper timing for appraisal. It is not possible to specify what the criteria should be as this will depend on the particular needs of the organization.

The specific purposes for which performance appraisal have been used (Bittner, 1948; Spriegel, 1962) can be grouped into two broad classes.

3.1 Employee Development

This has the general aim to help employees improve performance by discussing areas in which the employee needs improvement. Some of the specific purposes are:

(1) To discover workers' weaknesses as a basis for planned training.

(2) To help in assigning work in accordance with workers' ability.

(3) To stimulate people to improve.

(4) To develop people's morale through stimulating confidence in management's fairness.

(5) To provide an opportunity for supervisor and worker to discuss work problems.

(6) To get managers to look at their people and promote mutual understanding.

3.2 Administrative Action

Here the aim is to supply information to management to guide or justify administrative actions dealing with the employee. Some of the specific purposes are:

(1) To help in deciding who should be promoted, demoted, or given a raise in pay.

(2) To uncover exceptional talents.

(3) To furnish a basis for discharge of totally unfit employees.

(4) To serve as a check on employment procedures generally and interviews and tests specifically.

One cynic (Rieder, 1973) felt that performance appraisal was a plot to make the Personnel Department happy and to make them look good.

3.3 Cautions in Choosing Purposes

One rule, endorsed by virtually all the writers in the field, is to limit the purposes of the appraisal system. A major reason for the failure of appraisal systems is that they tried to do too much. Often the kind of information needed for administrative action (such as a ranking of employees) is counter-productive to employee development. Trying to do both with the same appraisal will lead to almost certain failure.

In industry, the two most common purposes of appraisal are some form of employee development and salary administration. Another rule, endorsed by virtually all writers, is to separate these purposes and handle each one by itself. The two purposes cannot be handled together. In trying to counsel an employee when he knows his salary hangs on a favorable evaluation, he will become defensive and blame everyone and everything besides himself for his shortcomings (Meyer, Kay and French, 1965). Hardly a conducive atmosphere to improve performance.

4. SHOULD THE WORKER BE TOLD?

The purpose of the appraisal system will determine, in part, whether the worker will be told how he was appraised. If the purpose of the system is to promote employee development, the worker must be told where he needs improvement (performance review). The method of telling the worker will be taken up later in this chapter. If the purpose of the appraisal is some administrative action, then, in theory, the worker need not know the details of his evaluation. In practice, however, provision for reporting back the evaluation to the worker is a necessity because workers want to know how they stand with their supervisor; and many of the possible benefits of the appraisal program cannot be achieved without it.

Bittner (1948) lists certain outcomes of an appraisal program which cannot be achieved unless the evaluation is reviewed with the worker:

1. Job performance can be improved by letting the worker know his weaknesses and strengths and making definite plans with him to overcome his defects and to make capital of his strengths.

2. Grievances can be prevented by letting the worker understand the basis for action which may be taken in the future and by clearing up misunderstandings about past actions that have affected him.

3. The supervisor and the worker can be brought into a closer personal relationship wherein each has a better understanding of the other, and the worker is made to feel that he is a person and not just a clock number.

4. Pent-up emotions which may be reflected in acts of aggression toward management may be relieved by providing opportunity for rebuttal and talking out the situation.

An additional outcome which can only occur from discussion is that the supervisor and the worker can define what is expected and what is considered important to proper job execution. For example, in a study done by Parker, et. al (1959) it was found that supervisors rated conscientiousness first and amount of work done second in importance. The worker reversed the order in what they thought was important. A similar result also reported by Prien and Liske (1962) and Barrett (1966) concluded that "typically incumbents and their supervisors disagree on how the incumbent does his work". These differences, however, can be worked out between the supervisor and incumbent if handled properly in a performance review.

Despite the potential benefits from discussing evaluations with employees, the practice is not universal (Spriegel, 1962). Although not all, most companies do discuss evaluations with employees.

4.1 How to Conduct Performance Reviews

It is generally agreed that the proper vehicle for discussing evaluations with the employee is a conference (or interview) between the rater (usually the supervisor) and employee. It is felt that the performance review or appraisal interview, as it is often called, is the single most critical factor in determining the success of the entire performance appraisal program (Burke and Wilcox, 1969; Meyer and Walker, 1961). To quote Meyer and Walker (1961), "The skill with which a supervisor handles the appraisal feedback discussion with his subordinate is a key factor in determining whether or not the performance appraisal program is effective in motivating

behavioral changes."

To be successful in motivating employees to improve performance a constructive atmosphere must be fostered in the discussion. If the employee becomes defensive or refuses to accept the process he cannot be expected to make the commitment necessary to improve his performance. Gibb (1965) concluded that defensive behavior is produced in an environment of evaluation, superiority, strategy, control, neutrality, or certainty. These qualities are often inherent in performance evaluation, making it ripe for producing defensive behavior. Great care must be taken by the supervisor in handling the performance review. This requires skills most supervisors do not have. Supervisors must be trained to effectively handle the interview situation (Miner, 1968). Such training is rarely given, which may account for the failure of many performance review systems. When it is given, performance review usually works.

It is difficult to delineate the "one best method" for conducting an appraisal interview. Although it is difficult, some writers have attempted to write "appraisal interview cookbooks" (e.g., Planty and Efferson, 1951; Leskover, 1967; Hoppock, 1961). These texts are of limited value and often present prescriptions which have no proven value, such as indicating that Friday is a bad day to hold appraisal interviews or suggesting that strong points be discussed before weak ones. Hillery and Wexley (1974) found that such order of discussion made no difference.

There has, however, been research aimed at determining critical variables or styles of interviewing which foster positive and negative results.

Maier (1958) describes three basic "styles" of interviewing. First is the "tell and sell" method. The goals of the method are to let the employee know how he is doing, gain his acceptance of the evaluation, and to get him to follow the plan outlined for his improvement. Meyer, Kay, and French (1965) document what can happen in such an interview, illustrating the major problems with the method; fostering defensive reactions, creating negative attitudes toward the program and not fostering improved performance. Meyer, Kay, and French found the following:

*Praise was more often related to general performance characteristics, while criticism was usually focused on specific performance items.

*Subordinates reacted defensively about 54% of the time when criticized.

*Constructive responses to criticism were rarely observed.

*The more criticism a man received, the more defensively he reacted.

*In a follow-up 10-12 weeks later, the area the employee identified as most criticized by his supervisor showed the least improvement.

*In only 60% of the cases did supervisors insure that specific work plans and goals were made in areas they felt the employee needed the most improvement.

In short, Meyer, Kay, and French found that the system was not working and in fact was more disruptive than helpful. Maier recognized these problems with the "tell and sell" method but felt that it might be effective with new or inexperienced workers. Hillery and Wexley (1974) found evidence to support Maier's belief. Trainees wanted a directed, tell and sell approach and were not satisfied with a more participative approach.

Maier delineates a second approach to interviewing he calls the "tell and listen" method. The goal is to communicate the evaluation to the employee and then let the employee respond. Its virtue is in its alleged cathartic value. It is very close to a clinical encounter. The supervisor is non-directive rather than refuting arguments raised by the employee. This is rarely used, or even advocated in industry, because it requires unusual clinical skill on the part of the interviewer. Rather than reduce frustration and aggression, it may increase it because the supervisor does not change his evaluation or even agree or substantiate it nor does he offer resolutions to any disagreements.

The third type of interview discussed by Maier, called "problem solving" is, in one form or another, the approach advocated by most writers in the area. The approach takes the supervisor out of the role as "judge or God" and places him in the role of helper. The key is mutual problem definition by supervisor and employee with the employee suggesting a plan of action to resolve the difficulty. The plan may involve help from the supervisor, a change of work assignment, procedure, etc. It is designed to stimulate thinking rather than simply supplying solutions. It affords an opportunity for upward communications. Progress on past goals is reviewed, solutions are sought for job-related problems, and new goals are established.

Several studies have reported positive results from a problem solving orientation. Meyer, Kay, and French (1965) report that such

a program has resulted in a more positive attitude on the part of the employee toward appraisal and his supervisor, and greater improvements in performance than under the tell and sell approach. Burke and Wilcox (1969) found more improvement in performance the closer the interview approximated the problem solving approach. Blake and Mouton (1961) report an increase in "team spirit" expressed by employee and subordinate when the problem solving approach was used. Using a role playing technique, Hanson et al (1963) found that a goals orientation in the interview resulted in greater satisfaction, less emotional tension, a sense of teamness, more comfort about job performance, and less resistance to supervisor's suggestions. Bassett and Meyer (1968) found that employees who prepared their own appraisal, discussed it with their manager and then came back two weeks later with specific performance goals and again discussed them with their manager felt better about the process, showed less defensiveness, and improved their performance more than did those evaluated under a tell and sell approach. Similar results were reported by Kirk (1965). There is always the danger, however, that supervisors will not continue the time consuming process of problem solving interviews. Dayal (1969), for example, reports a goal oriented appraisal system that worked well for the first year then the interviews became as mechanistic as any other system. This points out the need to constantly monitor the appraisal system in its operation to insure that the aims of the system are being fulfilled.

Besides adopting a problem-solving atmosphere there are a few other attributes which seem to promote an effective appraisal discussion.

Zander and Gyr (1955) report that employees have positive attitudes toward the appraisal if their foreman was seen as having the best interest of the men in mind and knew what he was talking about. Burke and Wilcox (1969) concur that employees who were more satisfied with the day to day supervisory performance of their supervisor were more likely to be satisfied with their performance in the appraisal discussion.

Solem (1960) reports that if the interviewer talked more than the interviewee, negative results may result. This may be just another example of the negative results which can occur in the tell and sell approach.

In summary, it seems desirable to discuss performance evaluations with employees if the discussion is handled properly. A problem-solving approach appears from the literature to be the most effective. Supervisors can be trained to carry out such discussions and seem to prefer it over the "God like" tell and sell approach.

Bittner (1948) concludes that failure to build the appraisal program in the light of the demands of the post-appraisal interview may result in its emasculation.

5. WHO SHOULD APPRAISE?

The decision regarding who should actually do the appraisal is extremely important to an organization. In the words of Klores (1966):

> ". . .The biases of raters will in large measure determine the future philosophy of the organization insofar as their biases are manifested in the characteristics of those who are promoted. For it is those ratees who most satisfy the biases of the raters who will rise to higher positions within the company and come to have increasing influence upon the philosophy of the organization."

There is some experimental evidence to confirm Klores' contention. Kirchner and Reisberg (1962) found that effective supervisors look for initiative, persistence, constructive action and planning in their evaluation of employees while ineffective supervisors look for following orders, tact, good team efforts, getting along with others and loyalty to the company. This also raises the problem that an outstanding employee will be lost to the company if his supervisor does not recognize his strengths.

The question of "who will appraise" has ramifications beyond company philosophy. Consideration must be given to the accuracy of the appraisal, its reliability, and the motivational impact on the appraiser and appraisee. Traditionally, the employees' immediate supervisor has been the chief appraiser of his performance, but recent writings have suggested other alternatives, such as peer ratings, self rating, and for supervisors, subordinate ratings. Each of these procedures will be discussed.

5.1 Appraisal by Supervisor

The most common procedure in industry is to have the supervisor appraise those under him. Apparently this is the person most employees want, and probably expect, to appraise them (Miner, 1968; VanZelst and Kerr, 1953). Supervisor ratings have been shown to have validity (Thornton, 1968) and reliability (Barrett, 1966) if care is taken to overcome common pitfalls in performance appraisal. This requires training the appraisers and using an acceptable appraisal form.

Supervisor appraisals are not without their problems. The specific problem, such as leniency, halo, and consideration of irrelevant variables, will be discussed later in this report. Although supervisor appraisals are prone to such contaminations, so are all the other schemes to some extent susceptible, e.g., peer, self, or subordinate ratings.

One question asked is: "What level of supervision should appraise?" The immediate supervisor or the next level of supervision up could perform the evaluations. It is generally agreed that the most important requirement for a rater is that he is familiar with the person he is evaluating and the job requirements of the incumbent's position. This makes the immediate supervisor the logical choice. This is supported by empirical evidence.

Whitla and Tirrell (1953) found that immediate supervisor ratings were more valid than ratings by higher level supervisors. Keep in mind that the variable here is not "level of supervision" but

rather "levels removed from the ratee". Prien and Liske (1962) report high correlation between first and second level supervisors (r = .60) in their appraisal of employees. This would infer that their evaluations were somewhat similar and interchangeable. Not so. Comparing the supervisors' evaluations with the employee's own self evaluation showed a higher relationship for the first level supervisor than for the second level supervisor. What this means is that first level supervisors appraise incumbents closer to the way the incumbents feel about themselves than do the second level supervisors. This has implications for the appraisal review; for an appraisal which is not consistent with one's own self appraisal can engender hostility and undermine the positive benefits of performance appraisal and review.

Berry, Nelson, and McNally (1966) found that, in the military, the agreement between supervisors in their evaluation of subordinates' performance is to some extent a function of the supervisor's rank in the organization. Supervisors of different rank do not agree, or, at least, each level of supervision imposes different values or views the subordinate from a different perspective. Very little has been done to determine which view of the incumbent is more accurate. A major problem with using second level supervisors as evaluators is that their major source of information about the incumbents is the first level supervisor. If the evaluation is going to be discussed with the incumbent, it is difficult for the second level supervisor to discuss the evaluation based on second-hand information. The first level supervisor cannot be asked to

discuss an evaluation he did not actually make and may not even agree with.

Before concluding that only the immediate supervisors should do the evaluation, consider two points. First, it is possible that if the second level supervisor reviews the evaluations given by the first level supervisor, the validity of the evaluation may increase. This may be due to the first level supervisor taking more interest, devoting more time and being more committed to the program knowing that his supervisor will be evaluating his evaluation performance.

Second, the second level supervisor may be in better position to discuss administration actions such as salary decisions with the incumbent. This leaves the counseling function for the immediate supervisor. He can carry out his goal of improving employee performance with the performance review without being encumbered with salary questions. As was noted in the last chapter, the goals of administrative action and employee development should be kept separate during the performance review. By giving salary decisions to second level supervision, this goal is achieved. In a survey of ratees, 20% felt that salary evaluations should be made by second level supervisors because first level supervisors who could write best got the most salary increases and promotions for their employees. This was felt to be unfair.

Even if immediate supervisors are delegated to appraise, it would be unwise to expect all immediate supervisors to be good evaluators. From the literature, it seems that the best, most effective super-

visors are the best evaluators (Bayroff, Haggerty, and Rundquist, 1954). They are less lenient in their evaluations and differentiate between their employees in terms of performance better than the less effective supervisors (Kirchner and Reisberg, 1962). This may, in part, be due to a more favorable attitude toward the appraisal process by the effective supervisors (Gruenfeld and Weissenberg, 1966).

The more intelligent the evaluator, the better he is able to differentiate the traits being evaluated (Stockford and Bissel, 1949). This would result in less halo and a better picture of the strengths and weaknesses of the employee.

One drawback to using supervisory evaluations is that it is seldom possible to find more than one supervisor with adequate first-hand information about an incumbent to evaluate him yet most people in the field strongly recommend pooling the evaluations for a person from several appraisers (Barrett, 1966; Bayroff, Haggerty, and Rundquist, 1954; Bittner, 1948; Patterson, 1922). Pooling should be done only if all the people are equally competent to make the evaluation of the person. In short, two heads are better than one only when both heads have something in them.

To get around this shortcoming and to seek new sources of information, peer or "buddy" ratings have been suggested and used.

5.2 Appraisal by Peers

Peer or "buddy" ratings have been used in the military but industry seems less willing to adopt the procedure despite its apparent success (Booker and Miller, 1966). There are at least

three reasons why peer (and/or subordinate) ratings may more accurately reflect the incumbent's real competence than will an evaluation made by his supervisor.

1. A man's peers (and subordinates) are usually in closer contact with what he does hour-by-hour and day-by-day than is his supervisor.

2. A man naturally tries to present only his best side to his supervisor, but his peers and subordinates see him as he is.

3. Using peers (and/or subordinates) as raters makes it possible to get a number of judgments, the average of which will be a more reliable measure than a single measure alone.

In comparing supervisor rating with peer ratings, results are equivocal. Springer (1953) reports low correlations between supervisor and peer ratings ($r = .39$ to $.25$) in an industrial situation. Booker and Miller (1966), on the other hand, report high correlations between peer and supervisor ratings in a military setting. The divergent results could be due to different rater-ratee populations or different rating scales and procedures. Some studies have reported higher reliability for supervisor rating (Springer, 1953; Kleiger and Musel, 1953). Others report high or acceptable reliabilities for peer ratings (Booker and Miller, 1966; Hollander, 1954; Haggerty, Johnson, and King, 1959).

One significant point on which the data is consistent is that peer ratings are perhaps the purest and best measure of leadership available (Brooker and Miller, 1966; Hollander, 1954; Wherry and Freyer, 1949; Roadman, 1964). Their use in promotion decisions might be valuable (Miner, 1968).

Although supervisor rating is preferred over peer ratings (Van Zelst and Kerr, 1953; Miner, 1968) only one reference could be found which reported widespread negativism about peer ratings (Bittner, 1948). One reason for this negative attitude is that employees do not like to feel that they are "cutting their buddy's throat". A recent article (Kaufman and Johnson, 1974) points out that such feelings are unnecessarily generated. The two most common techniques of obtaining peer ratings are the nomination procedure and the ranking procedure. The most common and preferred method is the nomination technique in which each person is to choose a specified number of peers who are "highest" on some dimension (e.g., leadership or promotability) and an equal number of persons who are "lowest" on that dimension. Employees object most strongly to placing their peers in the "lowest" category. What is usually done is, for each person in the group, the number of negative (lowest) nominations is subtracted from the number of positive (highest) nominations received to yield a score. Many people are not nominated into either category by any peers and therefore receive zero scores. What Kaufman and Johnson found was that the simple frequency of positive nominations was the most valid measure. Negative nominations added little and may have actually reduced the validity of the process. Apparently, negative nominations are contaminated by reactions of raters to anti-social behaviors of the ratee which are unrelated to job behavior. Simply then, there is good reason not to require peers to nominate "lowest" category persons. This

should reduce the "cut throat" complaint and make peer ratings more acceptable to employees.

5.3 Appraisal by Subordinates

This source of appraisal is only applicable in the case where the performance of the supervisor is of interest. It can be used to evaluate supervisors and management from the first level of supervision on up to the top. The same three advantages enumerated for peer ratings are equally applicable here.

Maloney and Hinricks (1959) report on a subordinate rating program used at Esso Research and Engineering Company. To insure anonymity for the respondents (apparently a critical factor when subordinates rate their supervisor), elaborate controls were instituted. Code numbers were used; at least four subordinates had to rate a supervisor; only department secretaries knew which supervisor went with the code numbers; answers to open-ended questions were paraphrased by personnel clerks. Although other people report that supervisors do not like subordinate ratings (Miner, 1968; Bittner, 1948), Maloney and Hinricks reported that 75% of the supervisors wanted to be evaluated again by their subordinates. One feature of the Esso plan may account for this unusual positive attitude; no one but the supervisor knew how he was rated, neither his supervisor nor the personnel department saw his ratings. This, in essence, took any threat out of the appraisal. The information was his to do with as he wished. Apparently many supervisors used it as a basis for changing their behavior. Twenty-five percent of

the subordinates said they saw lasting changes in their supervisor's behavior. Eighty-eight percent of the supervisors said they have tried to change their behavior. Sixty percent of the subordinates and supervisors said that productivity was favorably affected by the program.

Unfortunately, little else has been written about subordinate ratings as a tool to develop supervisory skills. It seems deserving of consideration, however.

5.4 Appraisal by Self

A relatively new suggestion is to have the incumbent appraise himself and discuss the appraisal with his supervisor. This interest in self-appraisal parallels, and, in fact, is part and parcel of the current emphasis on the problem-solving management by objectives approach to performance review. The use of self-ratings has been suggested for appraisal of supervisors as well as rank and file workers.

As was mentioned in Chapter 2 of this report, often the individual rates himself higher than his supervisor rates him. Unless the supervisor is trained to handle such situations, the performance review can be defeating to an employee and be counter-productive. It is interesting to note, however, that employees who tend to over-rate themselves are often judged by their superiors to be the least promotable employees (Thornton, 1968).

Bassett and Meyer (1968) found that self-appraisal, as an input into appraisal discussions, compared to more boss-centered appraisals, resulted in more satisfying appraisal interviews, less defensiveness

by the subordinate and greater improvements in subsequent on-the-job performance. They caution, however, that self-appraisal may be inappropriate for a new inexperienced worker or one who is highly dependent. This is supported by the work of Hillery and Wexley (1974).

In essence, the decision to use performance appraisal for employee development leads to the decisions to discuss the appraisal with the employee. Evidence suggests that the best approach for discussion is the problem-solving approach. Using the problem-solving approach requires some sort of self-appraisal. The choice is clear once the basic aims of the program are delineated. The use of self-appraisal does not preclude the use of supervisor, peer or subordinate ratings, however. In fact, Miner (1968) suggests that all types be used for appraising supervisor performance.

In summary, supervisor rating by the immediate supervisor should be part of any performance appraisal program. Employees expect it and they want to know what their supervisor thinks about them. Supervisor ratings are one of the best sources of information for administrative actions and such ratings get the supervisor to think about his main responsibility--his men. For employee development, self-rating should be included into the program. For supervisors, the use of subordinate ratings appears valuable. Peer ratings should be used only when information concerning leadership qualities for promotion is needed.

6. WHAT WILL BE EVALUATED?

This question seeks to determine what aspects of the employee and his performance will be the subject of the appraisal. Should the incumbent be evaluated on neatness of appearance, dependability or ability to get along with others? Barrett (1966) identified three broad classes of items which serve as bases for appraisal; an employee's personality (emotional make-up, character, intelligence); his performance (the way in which he goes about his work--effort, responsibility, planning); and his products (quantity and quality of whatever is produced).

A look at the history of industrial performance appraisal shows a distinct evolution from a dependence on personality trait ratings through an almost equal passion for evaluating observable behavior to the more current compromise position of evaluating personality traits using observable behaviors for definition.

The current discussion, really just a carryover of past discussions, is whether person-oriented traits such as ability to work with others, conscientiousness, or initiative should be included in an evaluation. (Everyone agrees that job-oriented traits, such as ability to do complicated jobs, ability to work with minimum supervision, or knowledge of work, should be a part of an evaluation). There are those who think person-oriented traits should be included (Kern, 1966; Kavanaugh, 1973; Labovitz, 1969; Klores, 1966). Klores expresses the sentiments admirably:

> The raters' personal subjective opinions will express themselves in some manner and it is better that they be expressed on traits known to be largely subjectively evaluated rather than be forced into expression in the rating of traits assumed to be objectively determined.

The fallacy of this argument is that including subjective person-oriented traits in an evaluation form does not eliminate subjectivity in the evaluation of the more objective job-oriented traits.

Others feel that the emphasis must be on objective job-oriented traits (Heier, 1970; Brumback and Vincent, 1970; Buel, 1962; Donovan, 1970; Miner, 1968). The reasons given are that subjectively evaluated person-oriented traits lack reliability (i.e., two independent raters often cannot agree on the evaluation of the same person), their definition is often ambiguous, the rater cannot distinguish between person-oriented traits, it is difficult to discuss with the employee evaluations on person-oriented traits and they are often not valid indicators of job performance.

One study specifically compared ratings on person and job-oriented traits (Taylor, Barrett, Parker, and Martens, 1958). They found that job traits had higher inter-rater reliability, that raters tended to be more lenient in their rating of person-oriented traits, and that person-oriented traits were correlated with a rating of overall performance.

The list of traits to be evaluated, whether person-oriented, job-oriented or both, should be determined from a thorough analysis of the jobs to be covered by the evaluation. What is required of a

person in doing a job must be known before we can measure whether he meets the requirements. When it is impractical to develop a separate rating scale for each job, the goal would be to pick out for a general appraisal scale the important requirements that are common to many jobs. As will be discussed in the next chapter, some writers have suggested rating techniques which should be specifically developed for one job or at least a family of related jobs.

The traits to be included should be selected on the basis of the following criteria (Bittner, 1948):

> Observability. Can the rater actually observe this trait in action? Is the worker's possession of this trait clearly evident to the rater in what the worker does?
>
> Universality. Is the trait under consideration an important characteristic in successful performance of all the jobs to be rated? It is unlikely, too, that the trait could even be observed when the job does not call it into play.
>
> Distinguishability. Is the trait under question clearly distinguishable as meaning something different from another trait with a different name? Do they overlap so much in meaning that ratings on the two would be nothing more than two ratings on the same basic characteristics?

More recently, attempts have been made to make person-oriented traits more objective by defining them in terms of behaviors or including behavioral descriptions for the scale points. An example would be the following definition of "initiative":

> Consider his success in going ahead with a job without being told every detail.

With the following behavior scale points:

(1) Needs constant supervision, will not go ahead without direction.

(2) Can do routine jobs without being told every detail.

(3) Once explained the job, he can be expected to complete it without further direction.

The technique for developing behavioral scale points will be discussed in the next chapter as well as reviewing the literature concerning their effectiveness. Suffice to say that this seems to be a reasonable compromise for including person-oriented traits while maintaining a semblance of objectivity in the evaluation scales.

7. WHAT TYPE OF RATING TECHNIQUE?

Before discussing the various types of rating techniques available, it would be valuable to review sources of bias and distortion found in rating techniques. Often specific rating techniques were developed explicitly to reduce a particular source of distortion. An understanding of these sources will aid in the evaluation and selection of a particular rating technique.

7.1 Sources of Distortion

7.1.1 Reliability vs. Validity

In previous chapters, the terms reliability and validity have been used without a specific definition. The terms are often confused or used imprecisely. Reliability, in terms of ratings, refers to the consistency with which individuals are rated. There are really two types of rater reliability, inter-and intra-rater reliability. Inter-rater reliability answers the question "how consistently do two independent raters rate a group of individuals?" This is assessed by correlating the ratings given a group of individuals by two raters. An inter-rater reliability coefficient can be computed for each item on a rating scale. The correlation indicates whether persons rated high or low by one rater were similarly rated by the second rater although the actual ratings may be different. For example, if rater A was a more lenient rater and rated everyone two points higher than did rater B, the inter-rater reliability would be perfect, i.e., 1.00, yet the actual ratings would be different for the two raters.

The second type of rater reliability, intra-rater reliability, answers the question: "How consistent is a rater rating the same individuals at two different times?" If a rater cannot agree with himself at two different times, the value of the entire rating process is suspect. Several problems exist with intra-rater reliability. First, the rater will often remember the rating he gave the first time and merely repeat the rating causing the reliability to be artificially inflated. Second, the ratee may have changed since the first rating and the fact that the second rating is different than the first may reflect true changes in the ratee, rather than unreliability. For these reasons, intra-rater reliability is not as often used as is inter-rater reliability.

Unfortunately, ratings in general are notoriously unreliable. It is not uncommon to find inter-rater reliability coefficients of .50 or less, which is considered low by any standard. Usually, we are forced to accept reliabilities around .70 yet these would be unacceptable for paper and pencil tests. Some of the sources of unreliability are the rating form, the raters, and the ratee. Rating forms that use ambiguous trait names or descriptions or require ratings on traits which are not observable lead to unreliable results. The raters may have different standards by which they judge personal biases or rating idiosyncrasies which would contribute to unreliability. The ratee may be inconsistent in this behavior or display different behaviors in the presence of the two raters. In such cases, the raters are in essence making their ratings based on different information about the ratee. Given all

these sources of unreliability, it is no wonder that when a rating scale demonstrates respectable reliability it is assumed that the scale must be measuring something in the ratee. The hope is that that "something" is what was intended to be measured.

Validity is an attempt to answer the question: "Does the rating scale measure what it is supposed to measure?" Unfortunately, this is extremely difficult to assess in the case of ratings because often they are the only source of data available. For example, suppose you wanted to determine if a supervisor's rating of initiative is really measuring an employee's initiative. How would you measure initiative without having someone rate the employee? If an inexpensive, effective way could be developed, why bother to rate in the first place? This, in essence, is the dilemma. So how then is validity assessed? One method is to use inter-rater reliability as a measure of validity, hence the confusion of the terms. The argument runs that if two independent people can agree on the rating of a group of individuals then that indicates that the quality they are rating is actually there. The argument has some merit but not much. For example, if two people are rating initiative, and both feel that employees who ask a lot of questions don't show initiative, their rating will be similar (high inter-rater reliability) but they would not be measuring initiative--instead--perhaps self-confidence or sociability.

A second technique is to correlate ratings to such criteria as quantity of production or number of promotions. The reasoning is that individuals rated high should be the better employees. Unfortunately,

the correlations are usually very low indicating not much relationship between the two.

In general, validity (other than by assessing inter-rater reliability) is rarely done or is done without sufficient controls to allow unambiguous interpretation.

7.1.2. Errors of Leniency

Errors of leniency occur when a rater is too lenient in this rating. He does not give low ratings even when they are deserved. (The reverse, an overly stringent, tough rater, is said to be exhibiting "negative leniency" errors). Leniency can be seen by observing the range of ratings given by a rater. Theoretically, there should be some individuals rated low, some high, and the majority rated in the middle-satisfactory range. If ratings all pile up on the positive end, leniency is suspected.

Leniency errors create two problems. First, because all the ratees pile up on one end of the scale, it is difficult to distinguish between them for purposes of administrative action. Workers are led to believe that their performance is satisfactory or even meritorious when in fact it is not. This can only lead to trouble later.

Second, leniency errors make it difficult to compare people rated by different raters. If a promotion is to be made and two people are being considered from different departments, the department whose supervisor is lenient in his ratings will display a higher rating, although he may not be as good as the other man. For example, Stockford and Bissel (1949) report leniency differences between supervisors so

large that the best employees in one department were rated lower than the worst employees in another department.

What are some of the contributing causes of leniency errors? As will be discussed, the structure of the rating scale and procedure contribute to leniency. In addition, there are several "social-personal" factors which may increase the likelihood of leniency.

Stone (1970) suggests that inner feelings of the rater may contribute to leniency. For example, an evaluator who feels insecure may tend to give high ratings in an unconscious attempt to make himself look good since he is usually responsible for training and motivating those under his supervision.

A second factor possibly contributing to leniency is the expectation of discussion of the rating with the ratee. Stockford and Bissell (1949) found that mean ratings of subordinates were 24 points higher (on a 100 point scale) when supervisors knew that their ratings would be discussed, rather than simply forwarded to the personnel department as usual. On the other hand, Creswell (1963) found that there was no less leniency when supervisors rated subordinates confidentially (no possibility of discussion) than when they were rating on a non-confidential form. The need for more research is indicated, but the expectation of discussion seems likely to influence leniency.

The degree of acquaintance between rater and ratee may also contribute to leniency. There is some evidence that acquaintance, up to a certain point, leads to more accurate ratings (Ferguson, 1949; Freeberg, 1969). Beyond this point, however, bias may become a factor. Bradshaw (1931) feels that point is when acquaintance turns

to friendship. Stockford and Bissell (1949) found ratings correlated .65 with length of acquaintance. A similar finding was reported by Knight (1923), adding additional support for the effect of acquaintance on leniency.

Perceived similarity of ratee to rater may increase leniency (Stone, 1970). Senger (1971) reported that supervisors rate subordinates with similar values higher than those with dissimilar values. Quinn (1969) reported little effect of similarity when it was measured in terms of biographical data such as educational level, marital status, military grade, etc. Here again more research is needed, but it may be that the critical similarities are those in the psychological-value realm rather than biographical similarity.

The magnitude of the consequences of the rating for the ratee might influence leniency. The greater the consequences of the rating, the greater the leniency error. Taylor and Wherry (1951) found that raters were more lenient in their ratings when they were for administrative purposes than when they were for research purposes alone.

It is possible that the retaliatory ability of the ratee may influence the rater. For example, if subordinates rate their supervisor, the supervisor may be more lenient in his ratings of them for fear of retaliation. There are, however, no studies which have investigated this possibility.

How can leniency errors be reduced? Probably the best and most effective method is to train the raters to recognize leniency errors in their own ratings, understand why leniency may be occurring, and

indicate the importance of reducing it. This is usually sufficient to reduce the problem. It is encouraging to note that Bayroff, Haggerty, and Rundquist (1954) found no difference in the validity of the ratings made by hard and easy raters. Whether this is universally true is a matter for conjecture.

7.1.3 Halo Error

Over fifty years ago, Thorndike (1920) pointed out that some raters have a tendency to rate an individual either high or low on many factors because the rater thinks the individual to be high or low on some specific factor. He called this tendency the "halo" effect, the result is that the various traits on a rating scale all intercorrelate higher with each other than would otherwise be the case. In essence, the rating scales are not measuring separate traits, but are all measuring one thing. Basically, the raters are not distinguishing between the traits. Bittner (1948) presents several actual illustrations of halo. In a study of a 12 trait rating scale applied to over 1,000 men in industry, an analysis showed that only two traits were really being measured. In a study of a 10 trait rating scale applied to 2,000 Army officers, it revealed that only three traits were being measured--namely, sense of duty, physical and mental endurance and ability, and leadership. It was found that four of the ten traits predicted the total score almost perfectly.

It is extremely doubtful that raters can distinguish more than five traits. The addition of other traits would probably only result in the duplication of the ratings given to the first 4 or 5 traits. It is interesting that Spriegel (1962) in a survey of over 567 firms found that only 7.5% of the firms using rating scales had four or less traits evaluated. Thirty-five percent used from 5 to 9 traits and 31% used 10 to 14 traits. Two percent actually used over 50 traits to evaluate employees. Firms using that number of traits are needlessly burdening executives who fill out and evaluate the forms. It is impossible for a rater to distinguish that many attributes in a ratee. Halo error would probably reduce the number of independent traits actually being evaluated to less than five.

A major cause of halo is an inadequate rating scale with too many traits and ambiguous trait names and definitions. Halo can most effectively be reduced by training the rater to understand it, recognize it, and see the need to reduce it. Training dealing with the meaning of the traits on the scale also would reduce halo. Selecting a rating procedure that is less prone to halo can help as well.

7.1.4 Rating Irrelevant Factors

It sometimes happens that raters are influenced--perhaps unknowingly--by various factors that are extraneous to whatever is being rated.

For example, McCormick and Tiffin (1974) report that in a steel mill the average rating decreased as the job level decreased. "Tinners", a high status job, were rated higher than "openers", a low status job. In theory there should be as many outstanding "openers" as there are

outstanding "tinners" yet the raters seemed to be rating the job rather than the man in the job. Klores (1966) reports similar results for professional personnel. In a similar fashion, job difficulty has been found to correlate with ratings (Svetlik, Prien, and Barrett, 1964). Workers on hard jobs are rated higher than workers on easy jobs, even though some of the hard-job workers are not doing their hard job as satisfactorily as some of the easy-job workers are doing their easy job.

Another contaminating factor which often creeps into rating is the length of service of the ratee (possibly correlated with acquaintance). Rothe (1949) found evidence that ratings correlated with length of service in 2 of 3 laundry services ($r = .62$ and $.87$). In the one situation in which it was not found to contaminate rating, the rater had had some training in industrial psychology. This underscores the need and value of training raters.

Other miscellaneous things can enter into ratings to contaminate them. For example, Spector (1954) found that if the ratee accepted the rater's suggestions on improving performance, the rater significantly increased his rating. In the study, however, the ratee accepted the suggestions, but did not act on the suggestions, his performance did not change, yet his rating did.

Ghiselli and Lodahl (1958) found that supervisors were rated lower by their supervisors if the ratee's group contained a member with higher supervisory ability than the ratee himself or if the group was autonomous rather than dependent. The first characteristic

has nothing whatever to do with the ability of the supervisor being evaluated. The second may be negatively correlated to ability. It has been said that the best supervisor is one who is not missed when he goes on vacation. Work group autonomy should be encouraged, not downrated.

The selection of a rating technique must take into account these sources of distortion, leniency, halo, and irrelevant factors. The various types of rating techniques will be discussed and relevant literature reviewed concerning strengths and weaknesses of each technique.

7.2 Rating Techniques

7.2.1 Essay

The first forms for evaluating employees asked the supervisors to write a short essay-type evaluation of the incumbent. In the pure form, this is rarely used today. The major problems were: that it was impossible to compare the evaluations of two people, even if the same evaluator wrote the evaluation; often the evaluations referred to irrelevant characteristics or were so general as to be useless; and, the quality of the evaluation depended more on the writing skills of the evaluator than on the performance of the incumbent.

Although these problems have not been totally resolved, essay evaluations are often combined with other types of rating techniques. (Spriegel, 1962). For example, space may be provided for the rater to give an illustration of the degree of performance described or the form may ask for comments to justify the rating given. The worth

of such hybrid systems is questionable. At best it forces the rater to consider the justification for his rating. Probably, in many cases, a rater could think of justifications for any rating he wishes to make. Unfortunately, there have been virtually no studies investigating what is done with the essay portion of ratings or whether their inclusion changes the characteristics of the rating significantly.

7.2.2 Graphic Rating Scales

The first graphic rating scale was described by Patterson in 1922. In its simplest form it consists of a trait name and descriptive adjectives spaced under the line. The following is an example of an item from a classic graphic rating scale:

DEPENDABILITY: _____

Unsatisfactory	Below Average	Average	Above Average	Outstanding

The rater could put a mark anywhere on the line inferring infinite discriminability between points. The originators of the scale recognized that such discriminability was not possible so they converted the scale into ten equal length segments and assigned a value of 1 to 10 to the rating corresponding to the segment in which the mark fell.

Patterson reported good inter-rater (r = .76 to .87) and intra-rater reliability (r = .75). He did concede that leniency was a problem but only for some raters. Other investigators have reported lower reliabilities (e.g., Taylor, Erwin, and Hastman, 1956) than Patterson did. Ryan (1945) criticized the scoring method proposed

by Patterson in which the individual item ratings are summed to get a total score. He felt that this assumes the traits being measured can compensate for one another. A person rated high on job knowledge but low on initiative will get the same total score as a person high on initiative but low on job knowledge. Whether these two people are really equal is debatable. One of the most devastating criticisms leveled was that halo effects occur with regularity.

Suggestions have been made to modify the basic graphic rating scale to reduce leniency and halo and increase reliability. One simple suggestion (Ryan, 1945), which has neither reduced halo or leniency or increased reliability, has simplified scoring. Rather than use a continuous scale, a multiple step scale is used. The rater is forced to choose one of the alternatives. The following would be the multiple step version of the simple graphic scale presented before:

DEPENDABILITY:

☐	☐	☐	☐	☐
Unsatisfactory	Below Average	Average	Above Average	Outstanding

Most multiple step scales use from 5 to 9 alternatives.

A second suggestion has been to modify the way the ratings are made. Rather than rate each person on all the traits before rating the second person, all the persons are rated on one trait before rating anyone on the next trait. It was thought that this would reduce halo. A variation of this is for each trait, to rate first the best, then the worst individual and then the second best, second worst, alternating until all individuals have been placed on the scale before

rating on the next scale. It was thought this might reduce leniency as well. Unfortunately, neither of these are any improvement over the standard technique in terms of halo, leniency or reliability (Taylor, Ewin, and Hastman, 1956).

Another line of modification, which has been widely accepted, is to supply more information about the trait being rated. Two parts of the scale can be modified, the trait name description, and the alternatives or anchors for the scale. Peters and McCormick (1966) found that reliability was higher when job task anchors were used rather than numerical anchors. Barrett, in a series of articles (Barrett, et al, 1958, Taylor, et al, 1958) investigated the impact of modifying the trait name by adding a definition and/or adding behavioral anchors. They hypothesized that supplying a description of the trait and behavioral anchors would be superior. This was not the case; however, using just the trait name with behavioral anchors proved to be the most reliable ($r = .67$ compared to .51 for the other conditions) and displayed the least amount of leniency and halo.

In recent years, there has been suggested (Smith and Kendall, 1963) a slight modification on the behavior anchor idea, the notion of "behavior expectations" as anchors. The anchors used represent, not actual observed behaviors, but inferences or predictions from observations. Raters are asked to decide whether a given behavior they have observed would lead them to expect behavior like that in the description instead of statements such as "shows interest in patients' description of symptoms". (The first scale was used to assess nurses' performance). The anchors consist of expectations such

as "if this nurse were admitting a patient who talks rapidly and continuously of her symptoms and past medical history, could she be expected to look interested and listen."

Smith and Kendall also detail a very elaborate system for obtaining and scaling the behavioral expectancies. A study by Borman and Vallon (1974) suggests that the procedure used to develop the scale may be more important than the scale itself. Briefly the procedures are as follows:

a. A sample of raters contributes behavioral examples representing low, average, and high performance on the job in question. The investigator attempts to exhaust the job performance domain by requesting examples to cover the content of the job as completely as possible.

b. The behavioral examples are clustered by content and dimensions of performance are named and defined.

c. Each member of a separate sample of raters rates each example in terms of the desirability of the behavior and sorts them into the dimension categories. Mean desirability rating and standard deviations for each item are computed as well as the frequency with which it was assigned to various categories.

d. The investigator decides which anchors are to be included for each dimension based on the criteria of low standard deviations and rater agreement in the sorting task.

e. The investigator retains those dimensions which he feels have a reasonable number of anchors meeting the criteria in (d) above.

As can be seen, raters are involved intimately in the construction of the scale, the anchors are supplied by them and scaled by them. This may, in fact, be a very effective way to train raters (Campbell, et al. 1973) and secure their cooperation and endorsement of the procedure, which may be more important than the specific format ultimately developed. The procedure is time consuming. Zedeck, et al (1974)

isolated 22 dimensions of nursing performance with each rater supplying three examples of each dimension, resulting in 420 different examples which had to be categorized and scaled for desirability--no small task by any standard.

Evaluating the effectiveness of behavioral expectation scales, unfortunately, has not shown it to be much of an improvement over simpler methods. Campbell, et al (1973) and Zedeck and Baker (1972) found low to moderate inter-rater reliability (r ranged from .24 to .55). Zedeck and Baker also found that halo still existed even with the behavioral expectation anchors. Two studies have specifically compared behavior expectation scales to the more traditional graphic rating scale. Borman and Vallon (1974) used graphic rating scales with definitions of the traits and numerical anchors as a comparison against the behavioral expectation scale. In this case, however, the raters did not participate in the construction of the scale and some of the behavior descriptions may have been out of date. Their results showed no difference in reliability (r = .56 to .61), leniency or halo (mean item intercorrelation = .73 for both scales). Burnaska and Hollmann (1974) compared behavioral expectation scales to graphic scales which used only trait name and adjectives as anchors. The results were essentially the same as found by Borman and Vallon -- no difference between the scales in terms of reliability (r = .78 to .81), leniency or halo.

It seems then that the road to improve the basic graphic rating scale has moved very little. In general, the use of behavioral

anchors (although not necessarily behavioral expectation anchors) does seem to improve the traditional graphic or multiple step rating scale, and should be used if possible. A particularly strong reason for including behavioral anchors is that it makes it much more meaningful for the ratee if his ratings are discussed with him. Counseling and employee development are best based on behavior rather than an ambiguous trait name.

A relatively new offshoot of the behavioral anchored rating scale is the mixed standard scale (Blanz and Ghiselli, 1972). It is designed to minimize halo and leniency and also permits an evaluation of the reliability with which each individual is rated, each scale rates, and each rater rates. With most rating procedures so far discussed, the rater is presented with descriptions of different degrees of goodness of performance for each of a number of separate traits pertaining to job performance, and he selects the one which best describes the person to be rated. In the mixed standard rating scale there are descriptions of three degrees of each trait to be rated, and the rater must respond to every description. He indicates whether he considers the ratee to be better than the description, to fit the description, or to be worse than the description. To reduce the possibility that the rater will form a clear picture of an order of merit set of descriptions for each characteristic being rated, the scales, and the order of the three statements in them, are mixed in random order -- hence the name mixed standard scale.

Preliminary findings reported by Blanz and Ghiselli seem very encouraging for reducing halo and leniency as well as identifying unreliable raters--an accomplishment no other rating technique can boast.

7.2.3 Critical Incident Checklist

In historical perspective, the critical incident technique suggested by Flanagan (1949) was the impetus for including behavior anchors on graphic rating scales. The critical incident checklist is developed by having raters list behaviors that represent various aspects of work behavior ranging from those that are desirable to those that are undesirable. A group of "experts" rates each item on the degree to which they are considered to indicate favorable or unfavorable behavior. Items in which the experts agreed are included in the scale. The scale value of the item (not presented to the rater) is usually the median rating given that item by the experts. Jurgensen (1949) has raised objections to the median and suggests an alternate procedure. The scale is presented as a checklist. The rater merely indicates those statements which are descriptive of the individual in question. The score is the sum of the scale value for the items checked by the rater as being descriptive.

Behavioral, or critical incident checklists, have several advantages; they are based on observable behavior, it is difficult for the rater to be lenient without knowing the scale values for each item, both rater and ratee have positive attitudes toward it (Clingenpeel, 1962), it has good reliability (Knauft, 1948) and is effective as a basis of employee counseling and development.

One major disadvantage is that usually separate checklists must be developed for each job which is very time consuming. Uhrbrock (1950), however, has scaled 724 general behavior statements which could be used to develop checklists quickly and efficiently.

A purer form of the critical incident methodology as an evaluation method is the Employee Performance Record described by Flanagan and Burns (1955). In essence it is a form of essay evaluation but critical behaviors (good and bad) are the substance of the evaluation. The procedure was developed at General Motors and involved giving the supervisor a small, specially prepared book in which to record good and bad incidents of behavior for each man under his supervision. The recordings were to be made daily as critical incidents were observed. Flanagan and Burns report that the daily recording took less than five minutes each day. The purpose of the record was to aid the supervisor in his discussion with the employee concerning his strengths and weaknesses, not as a basis of disciplinary action. Results of the first year showed 98,566 incidents of good behavior and only 7,670 of bad behavior recorded. A very small percentage had only ineffective incidents of performance on their record. About one-fourth of the employees had neither effective nor ineffective incidents recorded. During the first four years, the proportion of employees turning in suggestions increased from 10.8% to 21.9% and disciplinary actions were cut in half.

Whisler (1962) reports on another company's experience with the Employee Performance Record. He found wide disparity between

supervisors in the number of incidents recorded for their men. In some departments it was .5 per employee over a six month period. For one department the average was 17 per employee. The supervisor apparently made more use of on-the-spot discussions of the incidents rather than wait for the 6 month performance review. It was interesting to note that the union did not officially recognize the program, but they did use the records for settling grievance cases.

All in all then the use of critical incidents, systematically collected, seemed to have no negative effects and may have encouraged more performance-oriented discussions between supervisors and their subordinates.

7.2.4 Forced Choice

Forced choice scales are an offshoot of the critical incident checklist. The technique was originally devised to reduce leniency and halo errors. In format, the rater is presented with a series of two or more statements, really critical incidents, grouped together in blocks. The rater is asked to indicate which statement in each block is most descriptive of the person being rated (and in some cases which is least descriptive). There are various formats which can be used depending on the number of statements in each block, whether they are good or bad incidents, and whether the rater picks statements most like, or most and least like, the ratee. Berkshire and Highland (1953) after testing the various forced choice formats conclude that the best is to present four positive (good incidents) and ask the rater to select the two statements most descriptive of the

ratee.

The selection of the statements for each block is based on extensive preliminary research to determine the degree to which each statement is considered by raters generally to be "favorable" or "unfavorable" (favorability index) and the extent to which the statements, when used in a rating situation, tend to discriminate between above-average and below-average individuals (discrimination index). The statements are grouped so that the statements in each block have similar favorability indices, but differ in discriminability.

The reasoning is that a rater, who may attempt to rate a man higher than the man's true worth (leniency), has no way of knowing which of the statements to check to raise the man's rating as all the statements appear equally favorable.

In theory, the technique seems sound, but in practice it has fallen short of the mark. There is evidence that forced choice scales can be faked (Bass, 1957; Mais, 1951), they have been shown to lack validity (Kay, 1959), and they are notoriously unacceptable to the rater (Rogers 1960). Cozan (1955) reviewed ten studies which assessed the validity of forced choice rating forms and concluded that "forced choice does not give consistently higher validity than more traditional graphic rating forms. . . it does not justify scraping old performance appraisal systems in favor of the costly and technically complex forced choice methodology".

Lepkowski (1963) reported very high correlations between forced choice and graphic rating scales indicating that forced choice adds little to performance appraisal.

7.2.5 Ranking

The rating techniques so far discussed generally rated a man against an external, fixed standard of performance rather than the performance of other men. In graphic or critical incident rating formats, each man is rated against specific behaviors or a general conception of good and poor. With such scales, everyone can be rated good or be shown to display specific behaviors. With rankings (distinguished from ratings), the men in a group are arranged from best to worst. In this way someone must be ranked last even if his performance is satisfactory.

There are several techniques available for ranking people; the simplest is to have the rater merely start at the top or bottom and list the workers in descending or ascending order on the trait being rated. Usually, only one global overall performance trait is used rather than separate specific traits as is done with traditional rating schemes. A variation of this is to pick the best, then the worst, and alternate picking the next best, next worst until all the people have been ranked. The preferred method, although somewhat time consuming, is the paired comparison technique. All possible pairs of people are presented on slips of paper. For each pair the rater checks the one who is better. From this a complete ranking can be made of all the men. Lawshe, Kephart, and McCormick (1949) report that to rank 24 men using a paired comparison procedure (responding to 276 pairs) required about 30 minutes of the rater's time, and showed good reliability. The number of pairs required to rate men increases quickly. For example, to rank 30 people, as opposed

to 24, requires 435 pairs, as opposed to 276 pairs.

The major problem with ranking is that it is useless for employee development and can engender cut throat tactics between men in a group. Any help given a fellow worker may serve to raise his rank above your own. Thompson and Dalton (1970) warn that ranking systems are more likely to lead to performance decrement than to improvement. As for use in administration action, such as promotion, they can be of value. Two problems exist; however, first it is difficult to combine two sets of rankings unless there are some people common to the two sets. A person ranked first in one set of ranks may have been ranked tenth had he been included in the other set. Second, because of the time consuming nature of ranking, usually people are ranked on overall performance only. This does not give an accurate picture of what human resources, skills, and abilities are available in the work force.

7.2.6 Forced Distribution

Forced distribution, as distinguished from forced choice, is a ranking technique which sets up categories, usually seven, from poor to good. The rater places the names of the workers into the categories with the restriction that only a certain percentage of the people can go into each category. The percentages are developed so that there is a normal, bell shaped, distribution of people in the categories, few in the end categories and more in the middle. This "forces" the rater to distribute the workers according to a pre-selected scheme. Klores (1966) reports that forced distribution does not guarantee biasless ratings. Raters dislike using forced

distribution because it requires them to distribute workers contrary to their perceptions. Often ability is not "normally distributed" within a work group. To force a normal distribution would pervert one purpose of performance evaluation, i.e., to get an accurate picture of the workforce.

7.2.7 Field Review Method

The field review method of performance evaluation is not a separate method of appraisal in the sense that graphic scales are distinct from forced choice or force distribution. In fact, the field review method might make use of any of the rating and ranking procedures so far discussed. Wadsworth (1948) introduced the system in a detailed article. Basically, a personnel department representative meets with the supervisor to discuss the performance of each of his men. The personnel man then writes up the evaluations reducing the burden on the supervisor. A major purpose of the system is to provide feedback to the personnel department on their selection, placement, and training procedures. Based on the evaluation, the worker might be transferred, promoted, fired, or trained. Unfortunately, nowhere in the system is provision made to discuss the appraisal with the worker himself. Habbe (1953) discusses the procedure as used at Gimbel's Department Store. To illustrate the lack of communication with the employee consider the following quote:

> If a superior feels that the individual has been given a fair chance to improve and has failed to respond, a termination date usually is set without further ado.

The system can be expensive. Habbe reports it requires 3.5 man years in the personnel department to carry out the program. Each employee is "field checked" four times in three years. The appraisals are done on time, adequate time is given to them and no one is passed by. If the field review specialist is well trained he should be able to spot supervisors whose performance requirements make it difficult for talented people to get recognition and development and give them special consideration so that their talents are not wasted as a result of their accidental assignment to someone who does not recognize their worth (Barrett, 1966).

7.2.8 Management by Objectives

This appraisal technique is a natural extension of the problem-solving approach in the performance review discussions discussed in Chapter 4. Under management by objectives appraisal the ratee is evaluated on his program toward meeting specific objectives worked out between the rater and ratee in previous discussions. Usually, a form of self-rating is involved in which the ratee assesses his own progress toward the goals. In terms of employee development, management by objectives offers excellent potential. It can also be useful for administration actions, but the system was not devised for that purpose.

Management by objectives is not without its problems, however. Levinson (1970) feels that many management by objective programs are self-defeating in that they stress concrete, measurable goals and tend to ignore such things as customer service quality; they only offer a very limited range of possible goals, mostly those that

further the company's goals and often ignore the personal desires of the employees. Coleman (1965) concurs with Levinson. The problems can be overcome, however, and progress in employee development can be achieved. It is important that goals be developed which serve the needs of both company and employee.

Management by objectives makes comparisons between employees, working toward different goals, very difficult. It is not easy to determine the relative difficulty of goals or to always determine which goals are more important in terms of higher order company goals. In a good management by objectives program, goals should originate from the top with each successive level translating those goals into appropriate objectives for their level. In this way, all levels are pulling in the same direction.

It appears that an effective management by objectives program is the best system for employee development and counseling available. The other techniques discussed are effective for employee development only insofar as they embody the principles of management by objectives.

7.2.9 Job Sample

Performance evaluation by job sample is usually considered feasible for only lower level jobs. In such an appraisal, a worker would be evaluated on the product he produces. A lathe operator might be asked to produce a part and be judged on how close he came to the specifications; a mail sorter might be evaluated on speed and accuracy of sorting real mail or dummy mail made up specifically for the test. To some, this seems like the most natural technique of evaluation, and it is rarely challenged for production-type jobs

in which a product is produced.

Job samples, however, do have problems. Workers may do good or bad work because of the quality of the machine they work on or the quality of work done by others on the product before they did their part. It is not always possible to separate the performance of the man from his machine and the work materials. Often performance must be measured on subjective criteria such as smoothness, overall quality, neatness, etc. In such cases, graphic rating scales or critical incident rating procedures are used with all their problems. Several investigations have found that subjective ratings of objective criteria do not correlate with the objective criteria (Gaylord, et al, 1951; Stockford and Bissel, 1949; Paul, 1968) calling into question their adequacy for evaluating even objective performance.

Even if job samples are used for production jobs, their use for evaluating other positions may be limited. Managerial assessment by job sample has been suggested (Jaffee, 1966). Meyer (1970) reports on the validity of the in-basket test as a measure of managerial ability. Basically, the test is a simulation of a company. The manager is to assume he has just been hired to fill a position due to sudden vacancy. He must respond to mail, memos, etc., left in the "in-basket" of the previous manager. How he handles the problems is objectively scored and used as an assessment. The problem with any of these simulation role-playing assessment techniques is that they do not always tap important skills, but they do hold some promise as a supplemental evaluation procedure.

Evaluation of research performance has used job sample techniques, measuring such things as number of publications, number of patents, number of citations of an author's work, etc. Whitley and Frost (1971) assail such criteria as fostering a publish or perish atmosphere, promoting quantity rather than quality, and not really portraying accurately the research ability of the individual. They conclude "scientific performance is not a standard set of events which can be measured in a similar manner in different institutional settings". This underscores the need to tailor the performance appraisal criteria to the specific situation under consideration. A management by objectives approach seems most amenable to tailoring.

7.2.10 Summary

Table 1 lists the performance techniques discussed above and for each one rates its applicability for fulfilling the two major aims of performance appraisal--employee development and administrative action. The best technique for employee development is the management by objectives approach followed by critical incident and graphic rating scales using behavioral anchors. For administrative action, the field review method is considered the best followed by graphic rating forms.

The best system is probably a combination of techniques. Examples might include graphic rating scales with space for essay comments, critical incident checklist combined with graphic rating scales, or management by objectives combined with graphic rating

TABLE 1. APPLICABILITY OF VARIOUS APPRAISAL TECHNIQUES TO THE AIMS OF PERFORMANCE APPRAISAL

APPRAISAL TECHNIQUE	PURPOSE	
	EMPLOYEE DEVELOPMENT	ADMINISTRATIVE ACTION
ESSAY	Fair	Poor
GRAPHIC RATING SCALES	Fair to Good*	Good
CRITICAL INCIDENT CHECKLIST	Good	Poor
FORCED CHOICE	Poor	Fair
RANKING	Poor	Fair
FORCE DISTRIBUTION	Poor	Fair
FIELD REVIEW METHOD	Fair	Excellent
MANAGEMENT BY OBJECTIVES	Excellent	Fair
JOB SAMPLE	Fair	Fair

*The more behaviorally oriented--the better

scales. The decision must be made in light of the objectives of the program and the time and expense that can be invested.

8. SHOULD APPRAISERS BE TRAINED?

The answer to this question is an emphatic YES! Probably, the major cause of failure in performance appraisal systems is the lack of training given raters. Throughout this report the need for training has been stressed; writers implore their readers to take adequate time and train the appraisers in the importance of the program, how to rate, and how to feedback information to employees to foster development (Heier, 1970; Stockford & Bissel, 1949; Buel, 1962; Back and Horner, 1973, Bittner, 1948). Despite all the recommendations, little in the way of training is done in industry. Bittner (1948) reported that 78% of the personnel department representatives of Owens-Illinois Glass never, or infrequently, sit down with the rater to help him make out his ratings. Spriegel (1962) surveyed over 500 firms asking what methods were used to train raters. The following is a list of the four training methods cited and the number of companies who use each.

> Provide each rater with a manual explaining the program, but give no other special instructions - 72 firms.
>
> Hold a meeting of all raters to explain the program; may also provide a manual - 158 firms.
>
> Provide some practice in rating to give an appreciation of the standards; may also provide manual and/or hold meeting - 82 firms.
>
> Have individual consultation and review with raters who give out-of-line ratings - 175 firms.

It is the author's opinion that these four "training methods" are inadequate and do not truly meet the needs of a real appraisal training program.

Two studies carried out by the Army (Bittner, 1948) demonstrate that training increases the validity of the ratings given. Stockford and Bissel (1949) reported that training increased the reliability of the ratings given as well. Furthermore, training is believed to be the most effective method for reducing errors of leniency and halo. It is for these reasons that rater training should be an integral part of any performance appraisal system. It should be introduced concurrently with the introduction of the appraisal system and not after bad rating habits have been formed--or, what is worse, after resistance to the procedure has developed.

A good appraisal training program may require several training sessions and workshops. The three main areas which must be covered in the training program are:

(1) The value and importance of the program.

(2) How to make ratings.

(3) How to conduct the appraisal discussion with the ratee.

The first area on importance and value of the program is really a selling job. Commitment and active participation by top management will aid in selling the program. In one sense, if the raters do not "buy" the system, it will have little chance of meeting its objectives. It is for this reason that Kindall and Gatza (1963) do not recommend pushing the program on unwilling personnel.

The second area on how to rate is best handled by lecture and supervised practice. Bittner (1948) lists essential features of this area of training.

1. Instruction on the meaning of the characteristics, traits, or behaviors to be evaluated.

2. Instruction on the meaning of the points on the scale used.

3. Instruction on the avoidance of common pitfalls in rating such as:

   a. Lack of objectivity - basing ratings on supposition, guess work, emotional bias.

   b. Rating one trait in light of ratings on other traits.

   c. Ratings on the basis of a single dramatic incident.

   d. Rating on the basis of general impressions.

   e. Restricting the spread of ratings.

4. Supervised practice and discussion of practice ratings made.

5. Instruction on how to interpret the ratings.

6. Periodic refresher training.

The third area on how to discuss appraisals with the employees is most critical for employee development. Lecture role playing and video tape may be effective techniques to use here. Things that should be included are:

1. How to listen

2. How to set meaningful goals for employees

3. How to increase employee participation in the goal setting process.

4. How to measure progress toward goals.

5. How to constructively handle lack of progress toward goals and/or negative evaluations of employee progress.

6. How to take criticism without becoming defensive.

7. How to follow up on employee progress.

8. Demonstration, role playing and practice with group discussion of the process.

It cannot be stressed enough. An organization must be willing to invest the time and resources necessary to do a good job of training or the appraisal system is doomed to failure.

9. OTHER QUESTIONS OF IMPLEMENTATION

9.1 Who will be responsible for the program?

A survey conducted by Spriegel (1962) revealed that usually (82% of the firms) the personnel or industrial relations department is assigned responsibility for carrying out the program. Others assigned responsibility were a committee, the immediate supervisor, or vice-president-general manager. The person or department assigned responsibility for the program may not actually do the appraisal. It would be their function, however, to insure that it was done and done properly.

Rowe (1964) suggests providing an administrator of status and ability, commensurate with the importance of the program. This can aid greatly in insuring an efficient program supported by management. Heier (1970) suggests a development committee as the best way to get widespread support for the program and insure technical accuracy. Combining the development committee with Back and Horner's (1973) suggestion of involving immediate supervisors in all stages of development might be ideal. By including first line supervisors on the development committee the realities of the day-to-day work situation can be planned for in the system.

9.2 When will evaluations be made?

The major consideration in deciding how often to evaluate is practicality. If evaluations are made too frequently, raters may feel that they are being unduly burdened by the extra work. As a result, they may tend to race through them in a slipshod manner.

If the evaluations are not done frequently enough, the employee is deprived of a valuable source of feedback and direction for improving his performance.

Back in 1922, Patterson was recommending a three-month interval between ratings rather than monthly! Now writers suggest semi-annually rather than annually--how times have changed. Zander and Gyr (1955), however, found that monthly feedback sessions (not ratings) resulted in a positive shift in attitude relative to a six-month review procedure. Culbreth (1971) simply suggests appraising employees whenever the employee requires it. This, of course, implies that the employee is being evaluated constantly in order to determine when he needs evaluation.

The type of appraisal system and the goals of the system will in part determine the frequency of evaluation. For example, for employee development, more frequent evaluations may be required than for administrative action. A management by objectives approach would require irregular intervals matched to the goals set. In any case, evaluations should be required at least semi-annually in order to give employees a minimum amount of information about how their supervisor views their performance.

Special evaluations should also be made at critical times during the employee's service. A special rating, maybe after three months, should be made for a new employee or an old employee who is on a new job. A special evaluation should be made at the time an employee terminates to determine what type of employee is leaving. Perhaps a statement should be included indicating whether the supervisor would or would not rehire the employee. Perhaps special rating

should also be made at the time of transfer by the person's old supervisor.

9.3 Will the supervisors have time to carry out the program?

One consideration in setting up an appraisal program that is often overlooked is that of the demands the system makes on the rater's time (Bittner, 1948). It is all well and good to say that nothing is more important than taking this periodic inventory of our personnel assets and liabilities and that surely no one should begrudge the time spent on it. The fact remains, though, that appraisal is in competition with many other things for the rater's time, and the accomplishment of these other things has a more direct bearing in the rater's mind on his bread and butter. This presents a dilemma because an adequate appraisal system requires the rater to devote considerable time to it if the results are to be worthwhile.

Several ways in which the problem can be attacked have been enumerated by Bittner (1948). Have top management take an active interest in the program and give it unqualified support. Include in the program a systematic way of freeing the rater from his other duties for sufficient time in which to make out his ratings. Stagger the distribution of the rating forms so that the rater completes a few each week until he is finished. This staggering can, Bittner feels, become irksome to the rater. Bayroff, Haggerty, and Rundquist (1954), however, found that rating performed at the end of a long series of ratings was less valid then the earlier ratings suggesting that staggering might increase validity.

9.4 Where should the system be implemented?

Kindall and Gatza (1963) feel that the system should start at the top and work down. This insures that lower levels perceive that upper level management is participating and actively supporting the program. In a management by objectives approach, it is mandatory that goals be set from the top so objectives set at lower levels will be aimed at attaining the higher order goals.

Another alternative, and probably safer, is to set up the program in one division or group to evaluate it and convince other divisions of its value. Care should be taken to select a test site which has a high probability of success. Picking a "tough nut to crack" as the first test is foolhardy. Any new system will encounter difficulties by the very nature that it is new. Throwing up obstacles before the program has had a chance to adjust itself might be too much for a new system. After the "bugs" have been worked out, it has had a chance to mature and has had a few successes under its belt and it will be ready to tackle more difficult applications. It's important not to push the program on unwilling personnel or they will sabotage it one way or another.

10. RECOMMENDATIONS

The following recommendations for establishing a performance appraisal system are directed toward implementation in the Applied Sciences Department located at NAD Crane.

It is felt that the appraisal system should have as its principle aim--employee development. Because NAD Crane is a government installation, administrative actions are under the jurisdiction of the United States Civil Service Commission. To develop an appraisal system for the Applied Sciences Department designed for administrative actions would duplicate the function of already established Civil Service Systems.

Basically, it is recommended that a management by objectives performance review system be implemented for all employees, supervisors, and managers in the department. In addition, it is recommended that subordinate evaluations be made of all supervisors and managers.

10.1 Management by objective performance reviews

Moore (1967) strongly recommends the management by objectives (MBO) approach as the best vehicle on which to structure a performance review system for research and development personnel. An effective MBO program must start at the top with clear statements of department goals. These goals are then translated into objectives by each supervisor for his contribution toward the department goals. The employees under each supervisor set their objectives so as to contribute to the objectives of their area. In this way, at least in theory, everyone is working toward a common set of goals for the

good of the entire department.

Presently, an MBO program is in effect in the Applied Sciences Department, but it is not carried down to the individual employees in a systematic manner. As a preliminary step before instituting the recommended performance review system, the following recommendations are made.

RECOMMENDATION 1. Institute a review of the present MBO system in the Applied Sciences Department

This review should examine closely the objectives set, the manner in which they are set (e.g., who is consulted, what criteria are used), the manner in which they are communicated, how progress toward them is assessed, attitudes and opinions of supervisors and employees toward the objectives, etc. This review should point up any shortcomings with the present system which would have to be ironed out before a systematic review program is instituted.

For implementation of the program, the following recommendation is made.

RECOMMENDATION 2. Constitute (elect or appoint) a committee made up managers, supervisors, and employees to guide the implementation and follow-up of the program

The committee (possibly 6 or 8 people) would be given responsibility for developing specific procedures, informing people about the program, working out difficulties that may arise, and evaluating the effectiveness of the program. It is critical that this committee be composed of individuals respected by the employees and who, themselves, are committed to the program.

At first, it might be better to start the program with supervisors

who are willing to cooperate and give the program a chance.

RECOMMENDATION 3. Supervisors be trained to conduct MBO performance review sessions with their employees

The training of the supervisors is critical and should be carried out with care, devoting adequate time to the process. Supervisors must be trained to help employees develop realistic goals and how to assess progress toward the goals. A good training program may have to be spread out over several weeks using group discussion, work shop exercises, role playing and video tape procedures. After the program is implemented, additional training sessions might be needed to work out initial problems and reinforce prior principles taught. It is here that the greatest probability of failure occurs. In haste to implement a system, inadequate attention is devoted to training. The result is usually failure, or at best, a highly inefficient system.

It is not recommended that performance reviews be held at some fixed interval. Rather, it is recommended that the supervisor sit down with each of his employees at least once every six months and discuss past performance in relation to past goals, set future goals, work out plans for achieving the goals and assessing progress toward the goal, and set a date for the next review. The date will depend on the goals and objectives set, maybe in one month, maybe three, maybe six months.

RECOMMENDATION 4. An evaluation of the system be made during the first year of the program to assess its effectiveness and problems

The development committee would be responsible for seeing that the

evaluation be done; they might appoint others to design and carry out the actual evaluation, however. The evaluation should look at the attitudes of the supervisors and employees, evidence of positive and negative behavior changes, the manner in which the performance reviews are carried out by the supervisors, and the cost in terms of supervisor and employee time. The results of the evaluation would be used to modify the program as necessary.

10.2 Subordinate Evaluation of Supervisor

RECOMMENDATION 5. A form be developed for subordinates to appraise their supervisor

Appendix A contains an elaborate set of forms used by Maloney and Hinrichs (1959). It could serve as a model around which to develop a form more suitable for research and development supervisors.

RECOMMENDATION 6. The evaluation forms will be anonymously filled out by the subordinates

To insure anonymity, at least four subordinates must evaluate each supervisor. The data for each supervisor will be averaged by a clerk, comments from the subordinates will be paraphrased by the clerk to insure that idiosyscratic phrasings cannot be used to identify the subordinate. The supervisor will receive only the means and paraphrased summaries, the original forms will be destroyed. It is important that the employees be told about these safeguards during the orientation sessions necessary for introducing the procedure to the subordinates.

RECOMMENDATION 7. The supervisor and only the supervisor will receive the data

To insure that only the supervisor receives his data, code numbers

will be assigned each supervisor and used on the questionnaires. The code numbers will be known only to the department secretary. The clerk, who summarizes the data from each supervisor's subordinates, will put the summary in a sealed envelope. The department secretary will deliver the sealed envelope to each supervisor. The code numbers will then be destroyed, new ones would be generated each time an evaluation was to be done.

RECOMMENDATION 8. The supervisor be evaluated by his subordinates every six months

RECOMMENDATION 9. An evaluation of this appraisal system be carried out during the first year of operation

The evaluation should consist of interview and questionnaires designed to tap both employee and supervisor attitudes and opinions concerning the program. Employees should be asked if the supervisor has changed his behavior and whether they think it has been for the good or decrement of the department. Based on the evaluation, necessary changes can be made in the procedure, the form, or training given supervisors.

APPENDIX A. SUPERVISOR APPRAISAL FORM TO BE USED BY SUBORDINATES (MALONEY AND HINRICHS, 1959)

SUPERVISOR EVALUATION FORM (FOR USE BY SUBORDINATES)

SERIAL NO. OF SUPERVISOR YOU ARE RATING ___

HOW LONG HAVE YOU WORKED FOR HIM? ____

DATE _____

WHAT IS YOUR OWN LEVEL?

____NON-SUPERVISORY

____GROUP HEAD

____SECTION HEAD

____ASS'T DIRECTOR

CHECK LIST ON PERSONAL TRAITS

How well does each of the following words or phrases fit this man?	FITS VERY WELL 1	2	FITS FAIRLY WELL 3	4	DOESN'T FIT VERY WELL 5	DOESN'T FIT AT ALL 6
Good technical man	___	___	___	___	___	___
Tactful	___	___	___	___	___	___
Indecisive	___	___	___	___	___	___
Considerate	___	___	___	___	___	___
Unselfish	___	___	___	___	___	___
Good listener	___	___	___	___	___	___
Easy going	___	___	___	___	___	___
Scared of higher authority	___	___	___	___	___	___
Apple polisher	___	___	___	___	___	___
Good at handling people	___	___	___	___	___	___
Inexperienced	___	___	___	___	___	___
Puts things off	___	___	___	___	___	___

How well does each of the following words or phrases fit this man?	FITS VERY WELL 1	2	FITS FAIRLY WELL 3	4	DOESN'T FIT VERY WELL 5	DOESN'T FIT AT ALL 6
Regular guy	____	____	____	____	____	____
Plays favorites	____	____	____	____	____	____
Has confidence in his men	____	____	____	____	____	____
Good technical background	____	____	____	____	____	____
Honest	____	____	____	____	____	____
Stubborn	____	____	____	____	____	____
Too conservative	____	____	____	____	____	____
Sets good example	____	____	____	____	____	____
Immature	____	____	____	____	____	____
Helpful	____	____	____	____	____	____
Fair	____	____	____	____	____	____
Receptive to new ideas	____	____	____	____	____	____
Jumps to conclusions	____	____	____	____	____	____
Hard worker	____	____	____	____	____	____
Treats people like numbers	____	____	____	____	____	____
Not forceful enough	____	____	____	____	____	____
Doesn't know how to delegate	____	____	____	____	____	____
Overemphasizes petty details	____	____	____	____	____	____
Has the respect of his men	____	____	____	____	____	____
Technically competent	____	____	____	____	____	____

How well does each of the following words or phrases fit this man?	FITS VERY WELL 1		FITS FAIRLY WELL 2	3	DOESN'T FIT VERY WELL 4	5	DOESN'T FIT AT ALL 6
Lacks backbone	____		____	____	____	____	____
Aggressive	____		____	____	____	____	____
Does most of the talking	____		____	____	____	____	____
When you ask him a question, he gives you or gets you an answer	____		____	____	____	____	____
Wants his men to get ahead	____		____	____	____	____	____

CHECK LIST ON RESULTS

How would you rate the group(s) this man supervises?	TOPS 1	BETTER THAN MOST 2	ABOVE AVERAGE 3	AVERAGE 4	BELOW AVERAGE 5
On espirit de corps (team spirit)	____	____	____	____	____
On creativity	____	____	____	____	____
On importance of project assignment	____	____	____	____	____
On overall performance	____	____	____	____	____

CHECK LIST ON JOB METHODS

How satisfied are you with the way this man:	VERY DIS-SATISFIED 1	SOMEWHAT DIS-SATISFIED 2	REASONABLY SATISFIED 3	VERY SATISFIED 4
Assigns work projects and outlines what he wants done	____	____	____	____
Gives you room for individual initiative	____	____	____	____
Considers your personal wishes in making assignments	____	____	____	____
Listens to your ideas and suggestions and uses them	____	____	____	____
Trains and helps you do your job better	____	____	____	____
Keeps up to date on what you are doing	____	____	____	____
Lets you know when he has criticisms of your work	____	____	____	____
Lets you know when he thinks you have done a good job	____	____	____	____
Explains his criticisms and the changes he suggests	____	____	____	____
Gives you the technical help and advice you need	____	____	____	____
Lets you make the decisions you should make	____	____	____	____
Sees that your abilities are fully used	____	____	____	____
Lets you know what you need to do to get ahead	____	____	____	____
Admits his own errors	____	____	____	____

CHECK LIST ON JOB METHODS

How satisfied are you with the way this man:	VERY DIS-SATISFIED 1	SOMEWHAT DIS-SATISFIED 2	REASONABLY SATISFIED 3	VERY SATISFIED 4
Stimulates you to do good work	____	____	____	____
Keeps you informed on matters affecting you and your work	____	____	____	____
Plans and organizes the work of his unit	____	____	____	____
Stands up for you, when necessary, to higher management	____	____	____	____
Has authority to make the decisions you feel he should make	____	____	____	____
Makes you feel you are working with, rather than for, him	____	____	____	____
Is willing to sit down and help you with technical problems	____	____	____	____
Is able to sell his ideas to higher management	____	____	____	____
Is able to give you competent technical help	____	____	____	____
Makes prompt decisions affecting the output of his group	____	____	____	____
Keeps work from piling up on his desk for clearance	____	____	____	____
Sticks with decisions once they're made	____	____	____	____

SUMMARY EVALUATIONS

1. Overall what kind of a job would you say this man is doing? What do you think of the results he gets? The methods he uses?

---

2. Do you like working for him? Why? Or why not?

---

3. In what respect is he a good supervisor?

---

4. What are his main shortcomings?

---

5. What do you think he can do about these shortcomings?

BIBLIOGRAPHY

1. Back, K., and Horner, M. Successful schemes for management appraisal. Personnel Management, 1973, 5, 30-33.

2. Barrett, R. Rating Scale Content: I. Scale information and supervisory ratings. Personnel Psychology, 1958, 11, 333-346.

3. Barrett, R. The influence of the supervisor's requirements on ratings. Personnel Psychology, 1966, 19, 375-387.

4. Barrett, R. Performance Rating. Chicago: Science Research Associates, Inc., 1966.

5. Bass, M. Faking by sales applicants of a forced choice personality inventory. Journal of Applied Psychology, 1957, 44, 403-404.

6. Bassett, G. and Meyer, H. Performance appraisal based on self-review. Personnel Psychology, 1968, 21, 421-430.

7. Bayroff, A., Haggerty, H., and Rundquist, E. Validity of ratings as related to rating techniques and conditions. Personnel Psychology, 1954, 7, 93-113.

8. Benjamin, R. A survey of 130 merit-rating plans. Personnel, 1952, 29, 289-294.

9. Berkshire, J. and Highland, R., Forced choice performance rating--a methodological study. Personnel Psychology, 1953, 6, 355-378.

10. Berry, N., Nelson, P., and McNally, M. A note on supervisor ratings. Personnel Psychology, 1966, 19, 423-426.

11. Bittner, R. Developing an employee merit rating procedure. Personnel Psychology, 948, 1, 403-432.

12. Blake, R. and Mouton, J. Power people and performance reviews. Advanced Management, 1961, 26 (7 and 8), 13-17.

13. Blanz, F. and Ghiselli, E. The mixed standard scale: a new rating system. Personnel Psychology, 1972, 25, 185-199.

14. Booker, G. and Miller, R. A closer look at peer ratings. Personnel, 1966, 43, 42-47.

15. Borman, W. and Vallon, W., A view of what can happen when behavioral expectation scales are developed in one setting and used in another. Journal of Applied Psychology, 1974, 59, 197-201.

16. Bradshaw, F. Revising rating techniques, Personnel Journal, 1931, 10, 232-245.

17. Brumback, G. and Vincent, J., Jobs and appraisal of performance. Personnel Administration, 1970, 33(4), 26-30.

18. Buel, W., Items, scales, and raters: some suggestions and comments, Personnel Administration, 1962, 25(5), 15-20.

19. Burke, R. Why performance appraisal systems fail. Personnel Administration, 1972, 35(3), 32-40.

20. Burke, R. and Wilcox, D. Characteristics of effective employee performance review and development interviews, Personnel Psychology, 1969, 22, 291-305.

21. Burnaska, R. and Hollmann, T., An empirical comparison of the relative effects of rater response biases on three rating scale formats, Journal of Applied Psychology, 1974, 59, 307-312.

22. Campbell, J., Dunnette, M., Arvey, D., and Hellervik, L. The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 1973, 57, 15-22.

23. Clingenpeel, R., How employees feel about performance appraisal. Personnel, 1962, 39(3), 70-77.

24. Coleman, C. Avoiding the pitfalls in results-oriented appraisals. Personnel, 1965, 42(6), 24-33.

25. Cozan, L. Forced choice: better than other rating methods? Personnel, 1955, 36, 80-83.

26. Creswell, M. Effects of confidentiality on performance ratings of professional health personnel. Personnel Psychology, 1963 16, 385-393.

27. Culbreth, G. Appraisals that lead to better performance. Supervisory Management, 1971, 16(3), 8-10.

28. Dayal, I. Some issues in performance appraisal. Personnel Administration, 1969, 32, 27-30.

29. Donovan, J. In defense of subjective executive appraisal: a comment. Academy of Management Journal, 1970, 13, 351-353.

30. Fergusen, L. The value of acquaintance ratings in criterion research. Personnel Psychology, 1949, 2, 93-102.

31. Flanagan, J. Critical requirements: a new approach to employee evaluation. Personnel Psychology, 1949, 2, 419-427.

32. Flanagan, J. and Burns, R. The employee performance record: a new appraisal and development tool. Harvard Business Review, 1955, 33(5), 1-8.

33. Ford, G. Build a winning team with better appraisals. Supervisory Management, 1964, 9(12), 14-17.

34. Freeberg, N. Relevance of rater-ratee acquaintances in the validity and reliability of ratings. Journal of Applied Psychology, 1969, 53, 518-524.

35. Gaylord, R., Russel, E., Johnson, C., and Severin, D. The relationship of ratings to production records: an empirical study. Personnel Psychology, 1951, 4, 363-371.

36. Ghiselli, E. and Lodahl, T. The evaluation of foremen's performance in relation to the internal characteristics of their work groups. Personnel Psychology, 1958, 11, 179-188.

37. Gibb, J. Defensive Communication. A Review of General Semantics, 1965, 22, 221-229.

38. Glickman, A. Is performance appraisal practical? Personnel Administration, 1964, 27, 28-32.

39. Gruenfeld, L. and Weissberg, P., Supervisory characteristics and attitudes toward performance appraisals. Personnel Psychology, 1966, 19, 143-151.

40. Haggerty, H., Johnson, C., and King, S., Evaluation of mail order ratings on combat performance of officers, Personnel Psychology, 1959, 12, 597-605.

41. Hanson, P., Morton, R., and Rothaus, P. The fate of role stereotypes in two performance appraisal situations. Personnel Psychology, 1963, 16, 269-280.

42. Harris, C. and Heise, R. Tasks not traits-the key to better performance review. Personnel, 1964, 41(3), 60-64.

43. Hayden, R. Performance appraisal: a better way. Personnel Journal, 1973, 52, 606-613.

44. Heier, W. Implementing an appraisal by results program. Personnel, 1970, 47, 24-32.

45. Hillery, J. and Wexley, K. Participation effects in appraisal interviews conducted in a training situation. Journal of Applied Psychology, 1974, 59, 168-171.

46. Hobbe, S., Merit rating plus, Management Record, 1953, 15, 323-324.

47. Hollander, E. Buddy ratings: military research and industrial applications. Personnel Psychology, 1954, 7, 385-394.

48. Hoppock, R. Ground rules for appraisal interviewers. Personnel, 1961, 38(3), 31-34.

49. Jaffee, C. Managerial assessment: professional or managerial perogative? Personnel Journal, 1966, 45, 162-163.

50. Jurgenson, C. A fallacy in the use of median scale values in employee checklists, Journal of Applied Psychology, 1949, 33, 56-58.

51. Kaufman, G. and Johnson, J. Scaling peer ratings: an examination of the differential validities of positive and negative nominations, Journal of Applied Psychology, 1974, 59, 302-306.

52. Kavanaugh, M. Rejoinder to Brumback "The Content in Performance Appraisal: A Review", Personnel Psychology, 1973, 26, 163-166.

53. Kay, B., The use of critical incidents in a forced choice scale. Journal of Applied Psychology, 1959, 43, 269-270.

54. Kay, E., Meyer, H., and French, J., Effects of threat in a performance appraisal interview. Journal of Applied Psychology, 1965, 49, 311-317.

55. Kern, R., Appraisals and things. Personnel Journal, 1966, 45, 407-409.

56. Kindall, A., and Gatza, J. Positive program for performance appraisals, Havard Business Review, 1963, 41, 153-166.

57. Kirchner, W. Relationships between supervisory and subordinate ratings for technical personnel. Journal of Industrial Psychology, 1965, 3(3), 57-60.

58. Kirchner, W. and Reisberg, D. Differences between better and less effective supervisors in appraisal of subordinates. Personnel Psychology, 1962, 15, 295-302.

59. Kirk, E. Appraisal participation in performance interviews. Personnel Journal, 1965, 44, 22-25.

60. Kleiger, W. and Mosel, J. The effect of opportunity to observe and rater status on the reliability of performance ratings. Personnel Psychology, 1953, 6, 57-65.

61. Klores, M. Rater bias in forced distribution performance ratings. Personnel Psychology, 1966, 19, 411-421.

62. Knauft, E. Construction and use of weighted checklist rating scales for two industrial situations, Journal of Applied Psychology, 1948, 32, 63-70.

63. Knight, F. The effect of the acquaintance factor upon personal judgment. Journal of Educational Psychology, 1923, 14, 129-142.

64. Labovitz, G. In defense of subjective executive appraisal. Academy of Management Journal, 1969, 12, 293-307.

65. Lawshe, C., Kephart, N., and McCormick, E. The paired comparison techniques for rating performance of industrial employees, Journal of Applied Psychology, 1949, 33, 69-77.

66. Lepkowski, J. Development of a forced choice rating scale for engineer evaluation. Journal of Applied Psychology, 1963, 47, 87-88.

67. Lekovec, E. A guide for discussing the performance appraisal, Personnel Journal, 1967, 46, 150-152.

68. Levinson, H. Management by whose objectives? Harvard Business Review, 1970, 48, 125-134.

69. Maier, N. Three types of appraisal interviews. Personnel, 1958, 35, 27-40.

70. Mais, R. Feasibility of the classification inventory scored for self-confidence. Journal of Applied Psychology., 1951, 35, 172-174.

71. Maloney, P. and Hinrichs, J. A new tool for supervisory self-development. Personnel, 1959, 36, 46-53.

72. Mayfield, H. In defense of performance appraisal. Harvard Business Review, 1960, 38, 81-87.

73. McCormick, E. and Tiffin, J. Industrial Psychology 6th Edition, Englewood Cliffs: Prentice-Hall, 1974.

74. McGregor, D. An uneasy look at performance appraisal. Harvard Business Review, 1957, 35, 89ff.

75. Meyer, H. The validity of the in-basket test as a measure of managerial performance. Personnel Psychology, 1970, 23, 297-307.

76. Meyer, H., Kay, E., and French, J., Split roles in performance appraisal. Harvard Business Review, 1965, 43(1), 45-51.

77. Meyer, H. and Walker, W., Study of factors relating to the effectiveness of a performance appraisal program. Personnel Psychology, 1961, 14, 291-298.

78. Miner, J., Management by appraisal: a capsule review and current references, Business Horizons, 1968, 11, 83-94.

79. Moore, R. Appraisal at its apogee. Personnel Management 1967, 10(1), 61-72.

80. Parker, J., Taylor, E., Barrett, R., and Martens, L. Rating scale content: III. Relationship between supervisory and self-ratings. Personnel Psychology 1959, 12, 49-63.

81. Patterson, D. The Scott Company graphic rating scale. Journal of Personnel Research (now called Personnel Journal) 1922-23, 1, 361-376.

82. Paul, R. Employee performance appraisal: some empirical findings. Personnel Journal, 1968, 47(2), 109-114.

83. Peters, D. and McCormick, E. Comparative reliability of numerically anchored vs. job task anchored rating scales. Journal of Applied Psychology, 1966, 50, 92-6.

84. Planty, E. and Efferson, C., Counseling executives after merit rating or evaluation. Personnel, 1951, 384-396.

85. Prien, E. and Liske, R. Assessments of higher level personnel. III. Rating criteria: a comparative analysis of supervisory ratings and incumbent self-ratings of job performance, Personnel Psychology, 1962, 15, 187-194.

86. Quinn, J. Bias in performance appraisals. Personnel Administration. 1969, 32, 40ff.

87. Rieder, G. Performance review--a mixed bag. Harvard Business Review. 1973, 51(4), 61-67.

88. Roadman, H. An industrial use of peer ratings. Journal of Applied Psychology, 1964, 48, 211-214.

89. Rogers, B. The current status of the United States Air Force Officer effectiveness report, unpublished master's thesis, Florida State University, 1960.

90. Rothaus, P., Morton, R. and Hanson, P. Performance appraisal and psychological distance. Journal of Applied Psychology, 1965, 49, 48-54.

91. Rothe, H. The relation of merit ratings to length of service. Personnel Psychology, 1949, 2, 237-242.

92. Rowe, K. An appraisal of appraisals. Journal of Management Studies. 1964, 1, 1-25.

93. Ryan, T. Merit rating criticized. Personnel Journal. 1945, 24, 6-15.

94. Senger, J. Managers' perception of subordinates' competence as a function of personal value orientations, Academy of Management Journal, 1971, 14, 415-423.

95. Sloan, S. and Johnson, A. New content of performance appraisal. Harvard Business Review. 1968, 46, 14ff.

96. Smith, P. and Kendall, L. Retranslation of expectations; an approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

97. Solem, A. Some supervisory problems in appraisal interviewing. Personnel Administration. 1960, 23, 27ff.

98. Spector, A. Influences on merit ratings. Journal of Applied Psychology, 1954, 38, 393-396.

99. Spriegel, W. Company practices in appraisal of managerial performance. Personnel 1962, 39(3), 77-83.

100. Springer, D. Rating of candidates for promotion by co-workers and supervisors. Journal of Applied Psychology. 1953, 37, 347-351.

101. Stockford, L. and Bissel, H. Factors involved in establishing a merit rating scale. Personnel, 1949, 26(2), 94-116.

102. Stone, T. Sources of evaluator bias in performance appraisal. Experimental Publication System, October 1970, 8, Ms #290-12.

103. Svetlik, B., Prien, E., and Barrett, G. Relationships between job difficulty, employee's attitude toward his job and supervisory ratings of employee effectiveness. Journal of Applied Psychology, 1964, 48, 320-324.

104. Taylor, E., Barrett, R., Parker, J., and Martens, L. Rating scale content II. Effect of rating on individual scales. Personnel Psychology, 1958, 11, 519-533.

105. Taylor, E. and Hastman, R. Relation of format and administration to the characteristics of graphic rating scales. Personnel Psychology 1956, 9, 118-206.

106. Taylor, E. and Wherry, R. A study of leniency in two rating systems. Personnel Psychology 1951, 4, 39-47.

107. Thompson, D. Performance reviews: Management tools or management excuse. Personnel Journal, 1969, 48, 957-961.

108. Thompson, P. and Dalton, G., Performance appraisal: managers beware. Harvard Business Review 1970, 48, 149ff.

109. Thorndike, E. A constant error in psychological ratings. Journal of Applied Psychology, 1920, 4, 25-29.

110. Thornton, G. The relationship between supervisory and self-appraisal of executive performance. Personnel Psychology 1968, 21, 441-455.

111. Uhrbrock, R. Standardization of 724 rating scale statements. Personnel Psychology, 1950, 3, 285-316.

112. VanZelst, R. and Kerr, W. Workers attitude toward merit rating. Personnel Psychology 1953, 6, 159-172.

113. Wadsworth, G. The field review method of employee evaluation and internal placement, Personnel Journal, 1948, 27, 238-249.

114. Wherry, R. and Freyer, O. Buddy ratings: popularity contest or leadership criteria? Personnel Psychology 1949, 2, 147-159.

115. Whisler, T. An evaluation of performance appraisal in the G Company. In Whisler, T. and Harper, S., Performance Appraisal. New York: Holt, Rienehart, and Winson, 1962.

116. Whisler, T. and Harper, S. Performance Appraisal. New York, Holt, Rienehart, and Winson, 1962.

117. Whitla, D. and Tirrell, J. Validity of ratings of several levels of supervisors. Personnel Psychology, 1953, 6, 461-466.

118. Whitley, R. and Frost, R. The measurement of performance in research. Human Relations. 1971, 24(2), 161-178.

119. Wiley, L. Relation of Characteristics ratings to performance ratings. Journal of Industrial Psychology, 1964, 2, 7-15.

120. Zander, A. and Gyr, J. Changing attitudes toward a merit rating system. Personnel Psychology. 1955, 8, 429-448.

121. Zedeck, S. and Baker, H. Nursing performance as measured by behavioral expectation scales: a multitrait-multirater analysis. Organizational Behavior and Human Performance, 1972, 7, 457-466.

122. Zedeck, S., Imparato, N., Krausz, M., and Oleno, T. Development of behavioral anchored rating scales as a function of organizational level. Journal of Applied Psychology, 1974, 59, 249-252.

123. Zeitlein, L., Planning for a successful performance review program. Personnel Journal, 1969, 48, 957-961.